ARPA ORDER NO. 2223

ISI/RR-76-50
December 1976

Lawrence H. Miller

An Investigation of the Effects of Output Variability and Output Bandwidth on User Performance in an Interactive Computer System

INFORMATION SCIENCES INSTITUTE

UNIVERSITY OF SOUTHERN CALIFORNIA

4676 Admiralty Way/ Marina del Rey/ California 90291
(213) 822-1511

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ISI/RR-76-50	2. GOVT ACCESSION NO.
	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): An Investigation of the Effects of Output Variability and Output Bandwidth on User Performance in an Interactive Computer System.	5. TYPE OF REPORT & PERIOD COVERED: Research Report
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Lawrence H. Miller	8. CONTRACT OR GRANT NUMBER(s): DAHC 15-72-C-0308
9. PERFORMING ORGANIZATION NAME AND ADDRESS: USC/Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90291	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: ARPA Order [redacted]-2223
11. CONTROLLING OFFICE NAME AND ADDRESS: Defense Advanced Research Projects Agency, 1400 Wilson Blvd., Arlington, VA 22209	12. REPORT DATE: Dec[redacted]76
	13. NUMBER OF PAGES: 75
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): ----------	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document approved for public release and sale; distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This study is the author's Ph.D dissertation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Interactive computer system, man-machine interaction, output rate, output variability, statistical analysis, system output, user attitude, user performance

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(over)

DD FORM 1473, 1 JAN 73 — EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-014-6601

UNCLASSIFIED 407 952

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT

The performance of users in man-machine interaction (MMI) is described in terms of a number of user- and machine-oriented parameters. The general linear model of experimental design is used as a model of the interaction. Performance measures are selected and a questionnaire developed to gauge user attitudes toward the man-machine system (MMS) and its environment. The interface parameters selected are hypothesized to have a significant effect on the performance and attitude measures.

The effects of varying CRT display rates and output delays upon user performance and attitudes in a series of message retrieval tasks were evaluated experimentally. The results support the somewhat surprising conclusion that doubling the display rate from 1200 to 2400 baud produces no significant performance or attitude changes; increasing the variability of the output display rate produces both significantly decreased user performance and a poorer attitude towards system and interactive environment. The generally held notion that increasing output display rates is associated with better user performance is not supported; in fact, a general recommendation to system designers would be that increasing output display rates should not be attempted without a corresponding increase in CPU power.

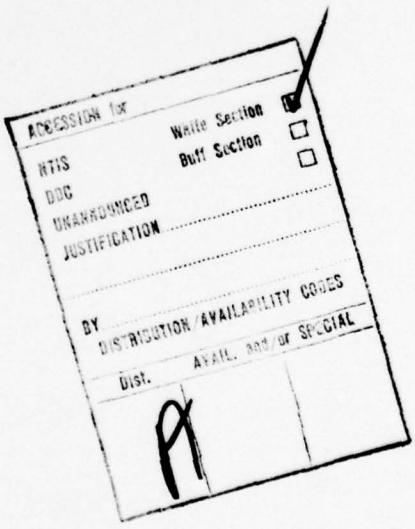
The questionnaire, which elicited user's attitudes toward the system, correlates with performance on the interactive tasks. The importance of these results to designers of MMS is discussed.

ARPA ORDER NO. 2223

ISI/RR-76-50
December 1976

Lawrence H. Miller

An Investigation of the Effects of Output Variability and Output Bandwidth on User Performance in an Interactive Computer System



INFORMATION SCIENCES INSTITUTE

UNIVERSITY OF SOUTHERN CALIFORNIA

4676 Admiralty Way/ Marina del Rey/ California 90291
(213) 822-1511

THIS RESEARCH IS SUPPORTED BY THE ADVANCED RESEARCH PROJECTS AGENCY UNDER CONTRACT NO. DAHC15 72 C 0308, ARPA ORDER NO. 2223, PROGRAM CODE NO. 3D30 AND 3P10.

CONTENTS

ABSTRACT

The performance of users in man-machine interaction (MMI) is described in terms of a number of user- and machine-oriented parameters. The general linear model of experimental design is used as a model of the interaction. Performance measures are selected and a questionnaire developed to gauge user attitudes toward the man-machine system (MMS) and its environment. The interface parameters selected are hypothesized to have a significant effect on the performance and attitude measures.

The effects of varying CRT display rates and output delays upon user performance and attitudes in a series of message retrieval tasks were evaluated experimentally. The results support the somewhat surprising conclusion that doubling the display rate from 1200 to 2400 baud produces *no* significant performance or attitude changes; increasing the variability of the output display rate produces both significantly decreased user performance and a poorer attitude towards system and interactive environment. The generally held notion that increasing output display rates is associated with better user performance is not supported; in fact, a general recommendation to system designers would be that increasing output display rates should not be attempted without a corresponding increase in CPU power.

The questionnaire, which elicited user's attitudes toward the system, correlates with performance on the interactive tasks. The importance of these results to designers of MMS is discussed.

This study is the author's Ph.D. dissertation.

I. INTRODUCTION

The general area of this research is man-machine interaction, specifically analysis of system output. Broadly stated, this study shows that the *variability* of system output bandwidth significantly affects users' performance across a wide range of interactive tasks.

In the field of computer science, it is now possible to divert attention away from fundamental theoretical issues towards refinements in systems and applications design for the greater satisfaction of end users. To this end, the previous ad hoc methods of refining systems to users' needs -- based on the intuition of the designer or programmer -- ought to give way to the more rigorous and reliable techniques of controlled observation, experimentation, and development. This research demonstrates that certain parameters of the man-machine interaction environment are manipulatable as a means of improving *user performance*.

This research uses an interactive message processing system now used at the University of Southern California's Information Sciences Institute and other locations. This program has been modified to provide a useful means of examining the relationship between the performance of the user and the variables influencing that performance.

Later sections discuss in greater detail the parameters of the interaction, which are shown to influence performance. Appropriate measurements are developed for evaluating performance in this kind of interactive task.

This work is an attempt to develop a model of man-machine interaction. Clearly a number of variables will affect performance in interactive tasks: for example, intellectual and cognitive differences in users, computer speed and power differences, command input differences, and data display differences. A complete model of the interaction would take all of these parameters into account, and use them to predict both user and system performance. Because this model was developed in order to predict user performance as a function of changes in the output parameters, it fixes the values of the other parameters to determine how changes in output affect user performance.

Statistical and experimental design [Winer, 1973] provide a framework for testing the validity of the model. The parameters hypothesized to influence performance are classified as independent variables and the performance measures as dependent variables. The objective is then to demonstrate that changes in the former produce changes in the latter. Statistical design theory suggests ways in which the problem can be structured such that the model may be tested and a probability statement made concerning the

likelihood that changes in the independent variables will actually produce changes in the dependent variables.

This research has involved two distinct phases. The first selected reasonable variables, which were hypothesized to influence user performance, and useful performance measures. The second tested for significance, i.e., the relationship between the independent variables and the performance measures. Ultimately, the rationale for performing research on the effects of changes in the system parameters upon user performance is to make possible the development of interactive computer systems that are more compatible with the needs and the limitations or abilities of the potential users of that system, as well as to provide a framework for testing future systems. In particular, Chapter 6 points out that controlled observation of people using an interactive system, under conditions which stress the functions available in the system, provides a means of identifying those aspects amenable to design improvement.

A number of broad-ranging questions occur as one studies the ways in which people interact with computer systems: suitable input language, keyboard and terminal designs, format and intensity of displays, amount of material, content of responses, speed and variability of display rates, etc. Additionally, there are questions concerning the way in which different individuals perform in the man-machine interaction: I.Q. differences, motivational and cognitive complexity factors, previous experience, etc. A complete theory of man-machine interaction (MMI) would take all of these factors into consideration in attempting to predict user performance for a given man-machine system (MMS). The theory would also include parameters relating directly to the individual system, and perhaps as well factors which relating to the supporting computer system -- CPU speed, memory capacity, etc. Any complete theory or model of user performance would, of course, have to predict performance measures that are useful; a preliminary step in developing a model or theory of MMI is the selection of useful or reasonable performance measures.

It is conceptually reasonable to break the set of MMI parameters into those which represent the man (user) and those which represent the machine. The machine parameters, in turn, may be divided into those which represent the particular interactive program, those which represent the interactive environment (the terminal, display and input language form), and those which represent the background processor.

By fixing all of the parameters except the display ones, this research explores the effects of changes in the display upon user performance in a given (though not untypical) MMS. In fixing the user population, the MMS, the input language form, and the background computer system, we still tacitly assume that the levels at which these have been fixed are representative of a broad class of interactive systems and users. It is this assumption of the generalizability of the research which makes the results of potential interest outside the immediate system and subject sample. Further elaboration on this concept of ***external validity*** is made in Chapters 4 and 6.

Specific hypotheses are tested in this research. By limiting the parameters, the basic hypothesis is that, for both the given user population from which the sample was taken and the particular interactive system employed, there are *significant* performance differences between groups of subjects receiving different levels of the display variables. A more detailed elaboration of the techniques, both the physical experimental environment and the statistical and design techniques, is presented in Chapters 3 and 4.

The experimental sessions involve three phases. First, subjects are given an introduction to the system. Next, they are given a series of tasks to be accomplished through the use of the system, i.e., questions to be answered by selecting and examining messages from a data base. Since the data base is rather large, the search and boolean operations available in the system must be used in order to reduce the number of messages that must be examined. Upon finishing all of the tasks, the subject completes a questionnaire on a number of features of the system. The statistical model for the analysis of the subject's responses is contained in Chapter 3.

Discussions of the controlled testing of man-computer interface issues are sparse in the computer science literature. Chapter 2 of this report reviews some of the pioneering work from which the ideas presented in this work germinated. Since experimental design and statistical evaluation of experimental results are not within the mainstream of computer science, Chapter 3 briefly develops the statistical techniques necessary for the experimental design, as well as the analysis of variance and multivariate techniques used to analyze the task and questionnaire data.

Chapter 4 describes the experimental design used to test the hypotheses, describes the interactive system and the experimental environment (the physical setting), and presents a sample subject scenario. Finally, Chapter 4 addresses the issues of internal and external validity (generalizability) of the results.

The results of the experimental session are presented and discussed in Chapter 5, including summary analysis of variance tables, graphical presentation of answers to the post-test questionnaire, and correlational findings. Chapter 6 extends the discussion begun in Chapter 5, drawing conclusions and making inferences about the utility of the types of experimentation conducted in this work. The shortcomings of the controlled experimentation and observation technique are specifically mentioned, and suggestions are made for further research into the relationships between the parameters of the MMI and the performance measures. Finally, the desirability of a theory of MMI is suggested, particularly the benefits to system designers of intensive observation of potential users to better understand those features of a given interactive system that best satisfy users' needs.

2. RELATED TOPICS AND AREAS OF STUDY

The actual number of controlled studies -- either of specific systems and their user population, or broader theoretical studies of MMI -- is extremely limited, although a number of authors express the opinion that these studies are needed. Willmorth [1972] states:

> Designing an information system for human use implies task analyses to determine the human actions to be performed, the decisions to be made, and the information required to be displayed to the human and expected from him, followed by the optimal design of the man/system interface. . .Time and effort must be devoted to designing a well-human-engineered system.

Willmorth goes on to note that there is virtually no verified human engineering data for software and suggests an experimental methodology for examining the relationships between various versions of on-line planning systems and a set of (unnamed) performance measures or characteristics. The paper serves more as a call for ideas or research rather than as a detailed statement of valid or useful performance measures.

Bennett [1972] concludes that "After a careful search of the major human factors and applied psychology journals. . .there is remarkably little evidence of research undertaken for the express purpose either of increasing our understanding of man-computer interaction or of providing information that will be useful in the development of systems that are optimally suited to user's needs." He identifies three areas that would benefit from human-engineering expertise: (1) conversational languages, (2) the effects of computer system characteristics on user behavior, and (3) the problem of describing, or modeling, man-computer interaction. Bennett's main concern is with the utility of the human-engineering attempt at model building and its lack of benefit to a systems designer. He feels that there is too great a distance "between the symbolic concepts and real-world data." He further reiterates his call for research by noting that early work with interactive facilities had computer efficiency as the paramount consideration, and by noting that "the experience that makes optimum usage patterns obvious to the designer rests on a computer-oriented lore unknown to people who are not computer professionals." His final remark is worth quoting in its entirety for its clear statement of the need of a discipline of MMI design:

> Because the theoretical basis for incorporating user problem-solving characteristics into analytical models is so rudimentary, the resulting user interface technology will take the form of procedural rules used by designers to guide their creative judgment. Indeed, the challenge for research is to transform the current art of design into an engineering discipline by

developing an agreement on ways for characterizing user tasks, for allocating interface resources to meet task requirements, and for evaluating user effectiveness in task performance.

Specific examples of research designs, methodology and results in the MMI literature include Walther and O'Neil [1974], who studied the effects of both user characteristics (for example, evaluative attitude [i.e., prior attitude towards computers], experience with on-line systems), and program and terminal characteristics (TTY vs. CRT, flexible vs. inflexible command recognizer). They found significant effects for terminal type and interface flexibility, as hypothesized, but there were often significant interactions with user experience or evaluative attitude. The utility of their work is that it submits to experimental verification their hypotheses concerning the relationships between the user and system variables of their study. Since some of their results were counter-intuitive, they add evidence that there is a need for carefully designed, well controlled experiments on the relationship between user and system characteristics and *user* performance. Their performance measures were limited only to time for task and syntax errors, and did not include user attitudes towards the interactive system or its interactive environment. The specific system of their study was an interactive text editor; subjects were required to find and correct a number of mistakes in a body of text.

Hansen [1976] examined differences in performance of groups of users in solving complex problems in on-line vs. batch environments. His study suffers from the difficulties inherent in performing research on users of interactive computer systems in that the extensions of the results to actual real-world environments is not always justified. However, he concludes his work by noting "it is not necessary to predict accurately and in detail in order to be useful. Man-machine research may be effective if it serves only to help the designer to organize his thinking about how [users] perform, to enable him to distinguish those variables which are likely to be important, and to design ad hoc experiments to answer specific questions."

Melnyk [1972] administered a post-test questionnaire in her study of the effects of limited ("frustrating") bibliographic search systems vs. a more open or free-form ("non-frustrating") search system. The questionnaire included items concerning keyboard design, keyboard ease of use, printing speed, etc.; a significant difference was found between the responses of those who experienced the "frustrating" system and those who used the "non-frustrating" system. There are some clear methodological difficulties in her study concerning the nature of the differences in the "frustrating" vs. "non-frustrating" systems, but her attempt to elicit user attitudes towards the system and its environment are significant. Unfortunately, there seems to be no concern for effects of subject differences between the two groups in her experiments. She reports a broad range of background and terminal experience among her subjects, but does not appear to incorporate the necessary experimental or statistical techniques to control for these differences. Since her work pioneers in considering the process of people using computer

systems, and concerns itself with user-oriented issues, it must be viewed in terms of its selection of performance and attitude measures as well as its methodology.

Others who have done experimental studies of the effects of interface parameters upon user performance in interactive systems include Carlisle [1974], who was also concerned with interface complexity, and Ting and Badre [1976], who were more concerned with interactive modes and their effectiveness in teaching and the subject's subjective judgment of the operational quality of the features provided. Importantly, the authors believe that "the overall judgment of the usefulness of the system was taken to be an indication of the success of the [man-machine] interaction."

The few researchers who have performed experimental research on interactive systems are concerned that the user's view of the system is as important to the success of a man-machine system as performance measures such as time to complete tasks, errors, cost, etc. In fact, there is as ample a volume of work on the nature of useful performance measures for evaluating interactive systems as there is a sparsity of actual studies evaluating systems. Sterling [1974] discusses the need for "humanizing" computerized information systems and the difficulty in deciding just what that term means. Martin, Carlisle and Treu [1973], in examining the man-machine interface in a number of interactive bibliographic systems, note that there is a lack of "knowledge about the blend of ingredients that produces a comfortable man-machine interface." Treu, in a later paper [1975], suggests experimentation in the effectiveness of interface languages, where the performance measure is a user-oriented one, i.e., the amount of mental work or "think time" spent by the user just before using a command.

Thus it is reasonable to optimize systems in terms of user-oriented performance measures. A questionnaire (see Appendix 3) was administered to the subjects of this study as a means of eliciting their attitudes toward the system as they had just experienced it. The post-test questionnaire data was further analyzed to determine whether differences in the versions of the system of this study are associated with differences in user attitudes towards the system. Chapters 5 and 6 contain detailed discussions of the questionnaire.

Questionnaires designed to elicit the subject's attitudes and opinions on system features were used by Heafner [1975] in his study of input language types for interactive message processing tasks, and by Heafner and Miller [1976] in their detailed study of the functions needed in a military automated message processing system.

The study of time and delays in interactive systems is presented by Miller [1968], who lists a number of interaction modes (from first log-on through requests for lengthy compilations) and discusses reasonable time delays for system response. The reasonableness of a delay is based upon user expectation and the concept of psychological closure. He does not discuss the possible effects on the user of continuous excessive or unanticipated delays in response over a period of time. The effects of

repeated delays upon the user's performance and attitudes form the foundation of the research reported here.

Seven, Boehm and Watson [1971] were also concerned with the effects of delays in interactive problem solving. Their study forced users to remain away from the terminal (locked out) for varying periods of time and discovered that total problem solving time was lower at some longer delay times than at lesser delays. One conclusion is that system delays *greater* than a certain amount can be useful (if they are *predictable*) in that they may free a user to engage in other productive activities. Clearly the effects of delays on user performance and attitude warrant further study.

The effects of the variability in the output display rate on the performance and attitudes of users in interactive computer tasks forms part of the foundation of the research reported here. There is a long tradition within experimental psychology of concern with the reaction time (RT) of subjects in various stimulus/response settings. The motivations for RT experiments varies from deep concern about the effects of fatigue and boredom upon drivers or pilots, etc., people whose work entails long periods of potentially extreme boredom occasionally mixed with sudden, usually unpredictable, moments when quick reactions are required, to concern about the underlying neurological processes by which we discriminate between differing stimuli and construct the appropriate response, as a means of gaining insight into cognitive functioning.

A number of individual studies involving varying stimulus and inter-stimulus times and the associated effects on RT have found that RT increases as the variability increases. Mackworth [1970], Mostofsky [1970] and Davies [1969] all present extensive surveys of the long history of experimental work in the variables (signal characteristics, task variables, subject variables and environmental variables) which affect user performance (generally measured in terms of response latency but also including other variables such as physiological measures of arousal, etc.) in RT and vigilance tasks. Some of the earliest results indicate that decrements in performance occur as the variability in the inter-signal arrival rate is increased, with the best overall performance being found with a regular series of events. Specifically, Mackworth [1970] reports on experiments which examined three levels of inter-signal variability and found that reaction time was shortest with the minimum variability. In the medium variability, the longest RT was found with those signals which followed the shortest inter-signal arrival. McCormack and Prysiazniuk [1961] used three levels of inter-signal interval variability. They also found that the shortest mean RT was found with the regular interval and the longest with the most irregular.

These authors indicate a theory of RT which involves the expectancy of the next signal on the part of the subject. It appears that subjects perform best when the next signal occurs at a time approximately equal to the mean inter-arrival rate of all previous signals. As the arrival of the next signal occurs significantly before or significantly later than this mean arrival rate of previous signals, the subject's arousal is decreased and response suffers.

It is this apparent decrease in performance directly associated with increased variability of inter-signal arrival rate that led to the conjectures studied in this research, that increasing the variability of the output display rate in MMI tasks would be associated with decreased performance and attitude of users of the interactive system. It is clear, however, that the operations of reading a number of messages and responding to their content involves a greater amount of information processing than merely reacting to a single stimulus and responding with a simple manual operation. It is this greater information content of the stimulus in the tasks reported here which might lead one to believe that a greater level of subject interest occurs which could counter the effects of the variability of the output. But it has been observed informally that during periods of heavy computer system use, when the output and response times of the system are quite variable, that user frustration and dissatisfaction do occur. The experiments reported here are an attempt at demonstrating in a formal manner the existence of this performance and attitude decrement as system response variability increases.

Suitable performance measures for evaluating user performance are mentioned in a number of the above references. Those used in this research are discussed in greater detail in later chapters. The above references suggested additional attitude measures to be incorporated. One additional questionnaire item (question 15; see Appendix 3) was suggested by Cooper [1973] who conjectured that a cold, hard cash expenditure might prove to be a useful means of judging the utility of an interactive system. His suggestion was modified somewhat for use in this research.

3. THE EXPERIMENTAL APPROACH

There is a need in the computer sciences for refinement in applications design in order to optimize system and user performance. Unfortunately, the well-established techniques that would allow designers to accomplish controlled research have not been accessible to them for a number of reasons: concern for "more pressing" design needs, lack of familiarity or experience with the techniques of experimental design, etc. As a consequence, one occasionally sees such statements as "Our system is easy and natural to use," in reports on interactive systems, input languages, etc. without the verification studies to defend them.

This work develops and parameterizes a model of man-machine interaction. People interact with machines in different ways, and with different styles. We would like to be able to predict the results of that interaction, with some probability of success, on the basis of values of certain parameters of the interaction. For example, how do differences in individual intelligence, cognitive style, training, specific experience with the system at hand, or educational background affect the interaction? How do machine differences affect performance? How do faster machines, higher input and output rates, and larger, more powerful machines affect performance? In fact, what kinds of performance measurements are useful in deciding upon the efficacy of a particular man-machine system ?

First we restrict our focus of attention to MMI where the *machine* is an interactive computer system. The man in the MMS will now be called the *user* of the facility. We conceive of him as approaching the system with the intent of solving a problem which could be a mathematical task or an information request, etc. We may further assume that the interaction is iterative in nature, i.e., through some initial interaction with the system, the user develops a partial solution to his problem, and this partial solution is used to converge upon a better solution to his problem.

We see the MMI as a joint effort between user and machine towards a final state objective. This objective need not be clearly perceived by the user before interaction begins, and may change during the course of the interaction. [See Figure 3-1.]

The figure is meant to imply that interaction moves towards an objective, and that changes in the perception of the objective are fed back to the system, which in turn may cause the objective to change.

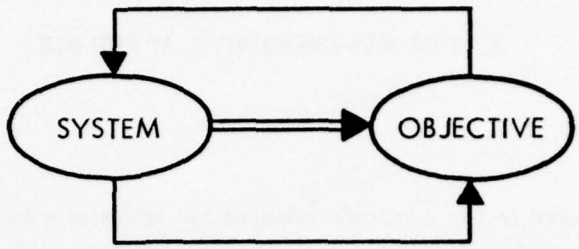


Figure 3-1. Schematic diagram of man-machine interaction

THE MODEL

We have a performance measure, P, which we would like to maximize in the MMI. P may be a set of measures $P = \{p_i \mid i=1,...,n\}$, containing values such as CPU time, memory, cost, time to complete the interaction, user satisfaction, frustration, etc. As described in the introduction, this work focuses on the user-oriented performance measurements described in detail below.

We describe P as a function of various parameters of the system:

$$P = f(p_1, p_2,...,p_n),$$

where the p_i's are values of the parameters, f a function of the p_i's. The p_i's relate to the man side of the MMI, and the machine side as well. Thus:

$$P = f(u_1, u_2,...,u_j ; m_1, m_2,...,m_k),$$

where $U = \{u_i\}$ is the set of user (man) parameters, and $M = \{m_j\}$ is the set of machine parameters. Some examples of these parameters have been mentioned above.

The function f may take a number of forms. One simple form, for which a great deal of analytical power is available, is a *linear* function. In those cases where it is believed that the relationship is in fact non-linear, suitable transformations may often be used to make the linear techniques appropriate. If f is a *linear* function of U and M, then

$$P = \sum_{\substack{i,j=0,...,\text{infinity}}} a_{ijkl} u_k^{\,i} m_l^{\,j} \quad k,l=1,...,kmax,\ lmax,\ \text{respectively}$$

P is a function of the u's, the m's, their higher order powers, and their cross products (interactions). The model is refined and simplified below.

This work is directed at selecting suitable u and m parameters, and testing the effect on performance of changes in these parameters.

A problem occurs: once we have selected u and m parameters that we believe affect performance, how do we demonstrate that this is in fact the case? Put another way, suppose we fix all values of $U = \{u_i\}$ and $M = \{m_j\}$, except one, some parameter p. We then observe a number of different people interacting with our system, each perhaps with a different value or level of p. We would expect performance to differ, and we take our measurements and observe differences. The problem is to decide whether these observed performance differences came about because of chance fluctuations, because of some changes in other factors which we did not keep constant, or as the result of changes in the parameter p. Additionally, the relationship between the parameter and the performance measure may be non-linear, so that the linear model is not sensitive to the relationship.

The performance that we *predict* as a function of this one variable, p, is conditioned upon the fixed values of the other U and M parameters. Thus $P = F(p \mid U, M)$. For the purpose of this research, as described in the introduction, we will assume fixed values of the u_i's. The value that we predict for P becomes a function of the machine parameters (M), conditioned upon fixing each of the user (U) parameters. Thus: $P = f(M \mid U)$, or to shorten, $P = f_U(M)$, where f_U is a function which depends on the specific values of the U parameters, and may be different for other values of the U parameters. This issue of the generalizability of the results is discussed in greater detail in Chapter 4.

We will attempt to construct f_U in such a way as to minimize the error in predicting P from the values of the parameters. The function is constructed from the set of observed pairs, (P_i, M_i), such that for each pair we calculate the predicted performance P'_i and an error, e_i.

$$P'_i = f_U(M_i) \qquad \textit{Predicted} \text{ performance}$$

$$P_i = f_U(M_i) + e_i, \qquad \textit{Actual} \text{ performance}$$

$$= P'_i + e_i$$

$$e_i = P_i' - P_i$$

Our criteria for constructing f include one that a function of $\{e_i\}$ should be minimized. The choices for this function of the errors include one such that the maximum error is minimized, over all possible f's as defined above. This "mini-max" solution leads to Chebyshev approximations. Since our data may contain "noise" or "outliers," this approximation is not useful here, since our choice for f would be inordinately influenced by highly random or noise effects. Another choice might be to minimize the absolute value of the errors, or the square of the errors. For consistency with the experimental

literature [Winer, 1971], our criterion will be to minimize the sum of the squares of the errors, or the least squares solution.

CONFIDENCE INTERVALS

It is necessary to know whether the solution represents merely chance fluctuations in the data, or is descriptive of underlying processes. Unfortunately, this question cannot be answered with certainty. The best that can be done is to assign a probability value or confidence interval to the solution. For example, in testing the performance of each of two groups in an MMI task, where the two groups receive different values or levels of one of the parameters, a confidence rating may be assigned to the observed performance differences. It might be possible to say that an observed difference as large as or larger than was observed would be expected to occur as the result of random fluctuations only, between two otherwise equal groups, only p per cent of the time. The levels of p at which we would be willing to agree that the observed difference is *not* the result of random fluctuations depends on the nature of the experiment, the cost we place on erroneous interpretations, etc., but historically a level of 1 percent or 5 percent has been used.

THE PARAMETERS

It is possible to select from a large number of parameters those which would be expected to affect user performance in an MMI. Previous research (see Chapter 2) indicates that those below may be expected to have a significant effect on performance. Furthermore, the variable "Output Variability" has implications for designers of time-shared interactive systems who are interested in allowing the largest possible number of users access to the system.

Independent Variables

(1) Output Variability (var)
Two levels: Low vs. High

(2) Mean Output Baud Rate (Baud)
Two levels: 1200 baud vs. 2400 baud

(3) Output Volume (Vol)
Two levels: < 1000 chars. vs. >1000 chars.

Dependent Variables

(1) Time to complete task
(2) CPU time
(3) Keystrokes used
(4) Post-test questionnaire (attitude survey)

These dependent variables will be loosely called "User Performance," or P.

Thus the model becomes:

$$P = f(v_i, i=1,...,n)$$

The *linear* model underlying multiple regression (or analysis of variance) implies that $P = L(v_i, i=1,...,n)$, where L is a linear function of the v_i's, or their products. If we use the three independent variables indicated above, each at two levels, the linear model reduces to:

$$P = a_0+a_1v+a_2V+a_3B+a_4vV+a_5vB+a_6VB+a_7vVB$$

where v represents the output rate variability, V the output volume and B the output baud rate. The statistical or experimental model used to test the *effects* of v and V on P will be a 2 x 2 x 2 factorial design (Figure 3-2 presents a simplified version).

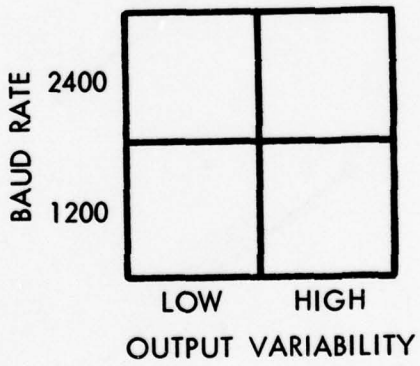


Figure 3-2. Factorial design: 2400 baud and 1200 baud vs. low and high variability

Repeated measures design was used as a means of reducing the between-subjects variability and to extract a maximum amount of useful information with a minimum number of subjects.

Figure 3-2 represents the possible combinations of system (independent) variables tested. Each "cell" of this factorial design represents one of the conditions, 1200 baud, low output variability; 1200 baud, high output variability; 2400 baud, low output variability and 2400 baud, high output variability.

Each subject in the experimental sessions is randomly assigned to one of the four conditions of the experiment. For each subject in each of the cells, a number of performance and attitude measures are taken as described above. The relationships studied in this research concern the performance and attitude differences between the subjects in the various cells of the factorial design. Specifically, it is conjectured that subjects experiencing the high variability versions of the system will show poorer performance and have a lower attitude towards the system than those experiencing the low output variability versions. Similarly, it is conjectured that those experiencing the 1200 baud versions of the system will have poorer performance and attitude towards the system than those experiencing the 2400 baud versions. A graph of performance vs. output variability for both the subjects experiencing the 2400 baud and those experiencing the 1200 baud versions would look like Figure 3-3.

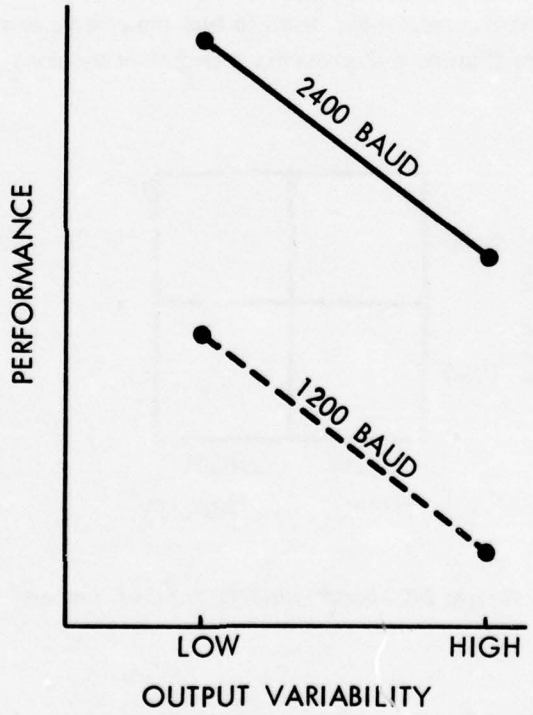


Figure 3-3. Hypothetical graph of performance vs. output variability for 1200 baud and 2400 baud versions

Classical analysis of variance [Winer, 1971] provides a framework and statistical model by which the separate and combined effects of the independent variables upon the performance measures may be tested. Specifically, given the observed differences in the cell means in the 2 x 2 x 2 factorial design, we attempt to determine whether these differences are the result of chance or random fluctuations in sampling where the

population means are actually identical, or are in fact different. Since our subjects are sampled from a larger population of interactive computer system users, the best we can do is make a probability statement concerning the observed differences, i.e., that observed differences as large as or larger than we actually obtained would be expected to occur *by chance* only, in otherwise equal populations, only p per cent of the time. The general linear model, of which classical analysis of variance is a special case, offers a procedure by which we may test the observed differences and determine the probability p described above. Historically, a level of 1 or 5 percent is selected as a priori levels for which we would be willing to accept or reject the hypothesis of no population differences.

The general linear model states that an observation or measurement upon subject k in cell i,j of an n x m factorial design can be described in terms of the separate effects of the overall mean of the measurement variable, the main effect for the first factor or independent variable, the main effect for the second factor or independent variable, the joint effect of the two of them combined, and an error term not accounted for by this linear model. Thus,

$$X_{ijk} = M + a_i + b_j + ab_{ij} + e_{ijk}$$

where X_{ijk} is the observation (measurement) of subject k in cell i,j; M is the overall mean of the dependent variable; a_i is the effect for variable A at level i; b_j is the effect for variable B at level j; ab_{ij} is the joint effect for A at level i and B at level j which cannot be accounted for by the separate effects of A and B (the interaction effect), and e_{ijk} is the error or residual in predicting X_{ijk} from the separate linear effects of A, B and the interaction. It represents all uncontrolled effects in the factorial design and may include higher order (non-linear) effects of the independent variables.

Classical analysis of variance provides the statistical tests necessary to test the separate effects for the independent variables and their interaction upon the performance measurement. We may define the within-cell variance to be the average variance within each of the cells of the factorial design:

$$MS_{within} = \Sigma (x_i - X)^2 / (n - 1)$$

where x_i is an observation from a given cell, X is the mean of that cell, and n is the number of observations in the cell. In this case, we assume that n is equal for each cell, and that the variances for each of the cells are equal. Since the linear model above defines the observation within a cell as a unique combination of the A effects, the B effects, the interaction effects and the uncontrolled or error effects, all observations within a cell would be equal if there were no error. Thus, we equate the within-cell variance, MS_{within}, to the error term in the linear model. We define the mean square due to the main effect for factor A by

$$MS_a = nq \Sigma (A_i \text{ (bar)} - M)^2 / (p - 1)$$

where n is the number of subjects per cell (assumed equal), q is the number of levels of the second independent variable (B), p is the number of levels of the first independent variable (A), A_i (bar) is the mean of A at level i, and M is the overall mean of the dependent variable. We define the MS due to the main effect for factor B analogously. Finally, the MS due to the interaction between A and B is defined by:

$$MS_{ab} = n \sum\sum (AB_{ij}(bar) - A_i(bar) - B_j(bar) + M)^2 \; / \; (p-1)(q-1)$$

where the terms are as defined before, AB_{ij}(bar) is the mean of cell i,j, and the double sum is taken over all of the p x q cells of the factorial design. In a 2 x 2 design, the denominator reduces to 1.

The hypothesis of no effects for any of the independent variables on the performance measure is equivalent to stating that the means of all the cells are equal. This is equivalent to stating that $MS_{between} = MS_{within}$, or equivalently, $MS_{between} / MS_{within} = 1.0$. When A and B are fixed factors (the levels of each in the factorial design represent all of the levels to which generalizations will be made), tests planned before the data are obtained may be made by use of the F statistic, $F = MS_{comparison} / MS_{within}$, with a sample distribution given by the F distribution at 1 and pq(n-1) degrees of freedom.

4. METHODOLOGY

The experimental design used to test the effects of the variables of this study on the performance measures was a 2 x 2 x 2 factorial design, with repeated measures on each subject across the two output volumes. A diagram of the factorial design is presented below (Figure 4-1). The levels of the independent variables selected for this study were

Output baud rate	1200	2400
Output rate variability	Low	High
Output volume	<1000 chars.	>1000 chars.

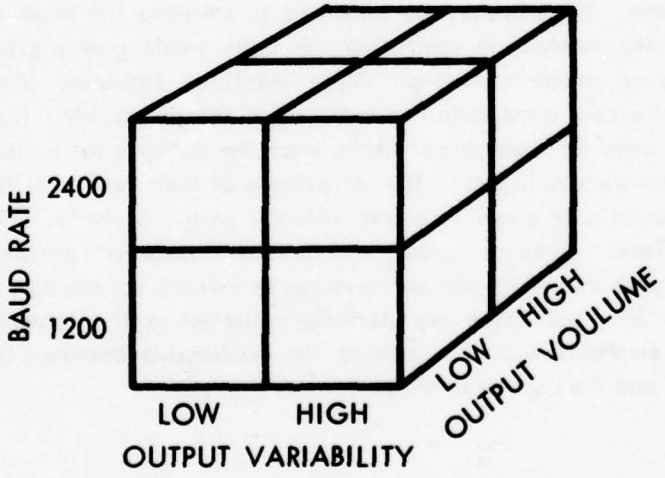


Figure 4-1. 2 x 2 x 2 factorial design

The performance measures used in this research have been described in the previous chapter. It is not initially clear just where the focus should be in order to improve the performance of an MMS. For example, tradeoffs are required whenever it becomes necessary to optimize one part of an interactive system. In general, increasing the output display rate requires either a more powerful processor or a limit on the number of users who may simultaneously use the interactive system. Beyond a limit in

the display rate, further attempts to increase the speed of output, with a given CPU and a given number of users, leads to noticeable degradation in system performance; in particular, *total* time to solve a problem, perform a compilation, etc., is increased. Even though the nominal output rate is increased, the *actual* output rate is decreased, since the processor can in general service only one terminal at a time, then must service others before returning. This leads to a high variability in the output display rate, with characters being displayed in bursts of variable length, and with a variable time delay between bursts.

One of the conjectures tested in this research is that this variability in the display rate is associated with a decrease in *user* performance in interactive problem-solving tasks, i.e, even when the nominal and actual display rates are similar, the variability alone would lead to reduced performance. Furthermore, it was conjectured that people who use a heavily loaded system (i.e., one with a great deal of variability in output display rate) would have a poorer overall view of the entire MMS and its environment. The post-test questionnaire represented an attempt to measure subjects' attitudes towards the entire MMS as they had just experienced it. Detailed analysis of the questionnaire is presented in the next chapter.

The selection of performance measures used in this research followed from the above considerations. In particular, the total time to complete the tasks and the number of functions used (keystrokes) in completing the tasks would give a gross indication of user performance in problem solving. Other measures reflecting user performance included number of errors made, help requests, use of references, etc. Number of errors, however, was not used for these experiments, since the subjects were instructed to work on a task until they were satisfied of the correctness of their response; thus the number of wrong answers on any given task was virtually zero. Similarly, CPU time used in completing the tasks gives a gross measure of computer performance. Other machine-oriented measures might include measures of memory access, disk access, memory management time, etc., but these are partially reflected in the total CPU time used. Furthermore, the emphasis was on examining the relationship between the independent (display) variables and the user-oriented performance measures.

THE SYSTEM

The system used to test the influence of the independent variables on the performance measures is an interactive message retrieval system in use at the Information Sciences Institute and other locations. It works on unstructured but formatted text files which conform to a standard message format. The program is regularly used by a number of users of the ARPA computer network and is routinely used by all of the subjects of these experiments. The subjects required essentially no additional training. This system has been modified to permit performance measurements to be taken on-line. Further modifications were made to the program in order to mold it to the simulated travel department environment of this study.

The data base consisted of approximately 200 travel request messages (see Appendix 4 for examples of messages from the data base). The messages were generated by a SAIL program [Smith, 1975; VanLehn, 1973] using a random number generator to select names of travellers, dates of travel and return, "fellow travellers" and destination cities.

Each message consisted of the following fields:

To: All were to the Travel Department

From: The name of the individual requesting the travel. In all cases, the person sending the message was requesting travel for himself, and up to three additional people.

Subject: Consisted of the word "Travel," followed by the destination city and the date of intended departure. This field was accessed as the "Destination or Date" field. It corresponds to the "Subject" field in the standard message format.

Message body: All messages were worded as follows:

Please reserve a seat [or N seats if more than one person travelling, where N is the total number travelling] to <destination> on <date of requested travel> for me [if more than one,

Second traveller
Third traveller
Fourth traveller]

RETURN: <date of return, or OPEN>
Thanks

The program consisted of the Main, MSG and Questionnaire modules.

The Main module initiates the experimental session. The identification of the subject is entered, and the values of the parameters for the subject are generated. The system returns to this module after all tasks are completed.

The MSG module is the modified message processing system. Subjects are instructed to assume that they are a clerk in the travel department of an organization. The system has been modified so that commands to the system relate to the types of requests which might be made of a data base of travel messages. These include commands that allow the user to search on Date of requested travel, Destination city or Name of traveller. The system allows complex boolean search requests combining parts of the message normally available for searching.

The usual result of a search is a listing of just the headers of the messages satisfying the search request. From these headers, the user may specify the exact message or messages to be displayed on the screen. If the user has reason to believe

that the selected messages will not be too numerous, or knows that he will want to read all of the selected messages, he may have the system immediately begin typing them on the screen rather than first displaying the headers.

The command structure uses single-letter commands. For example, to see those headers of messages where the requested date of travel was in April, the user would type

H D April

The system would echo back what he had typed by completing the command, and the user would actually see the following on his screen:

Headers Destination or date string: April

After a period of time during which the system is searching the entire data base and selecting just those messages which have the word April (case-insensitive) in the Date of Travel string, the headers of those messages would be typed out.

The Questionnaire module administers the post-test questionnaire to the subject (see Appendix 3), maintains the file of answers, and includes the screen editor for the final open-ended question.

THE SUBJECTS

The subjects used for these experiments were members of the professional and secretarial staff of ISI. In order to minimize training time and to better represent a population of experienced interactive computer users, subjects were taken from those who had already had some experience in using the system. Though most users of interactive computer systems do require training and experience with a particular system in order to reach maximum efficiency, the learning and adaptive phase represents only a very small fraction of the total time in which they will be interacting with the system. To use subjects who have not reached full familiarity with the particular test system leads to the problem of confounding learning effects with performance effects.

Subjects for this study represent a population of experienced interactive message search and retrieval system users. Though this population has been sampled only within the confines of a research institute environment, the broad cross section of subjects for this study -- male and female, professional and secretarial -- leads to the conclusion that this sample is representative of the types of people who might use interactive search and retrieval systems in other environments. Such individuals might include, but are not limited to, librarians or other users of interactive bibliographic systems, airline reservations personnel, hotel reservations service personnel, users of data base management or retrieval systems, etc.

Subjects were randomly selected from the entire ISI staff. The only requirement for participation (other than volunteering) was that the subject use MSG, the message processing system modified for this study, on a regular basis. The factorial design necessitated approximately ten subjects in each of the four cells of the design. Approximately 40 subjects were needed for the experiment. Due to system difficulties, however, the number of actual subjects was 9 per cell, a total of 36. Each subject was assigned to one of the four cells by a random number assignment procedure internal to the initialization routines. After approximately ten subjects, an attempt was made to weight the cell assignment so that the number of subjects per cell would be equal. Most importantly, however, subjects were assigned to baud and output variability conditions without regard to whether they were male or female, members of the secretarial or professional staff, without regard to time of day or current system loading. Some attempt was made, however, to ensure that no one cell contained all of the females, or all of the professional staff, etc.

PILOT STUDY

A pilot study was conducted prior to the main research effort reported here. It allowed for an initial testing of the relationship between the machine-oriented independent variables and the performance measures. It demonstrated the desirability of including output baud rate as an additional display variable, both because of its intrinsic potential effect upon user performance and also because of the potential confounding effect between output variability and actual display rate.

The pilot study afforded an excellent opportunity to refine both the experimental design and the physical design of the sessions. In fact, a "pre" pilot test was conducted using a small subject sample in order to iron out problems that subjects might have in the use of the Travel Message System. The experience gained in the pilot study made the main experimental sessions relatively free of the need for experimenter intervention. This was not entirely the case during the pilot study.

Data analyses were accomplished on the data taken during the pilot study, and they made it possible to refine actual statistical procedures accomplished on the main experiment data. The results of the analyses of the pilot study data strongly agree with the results obtained in the main study. The pilot study provided initial support for the validity of the conjectured relationships between the display variables and the performance measures.

EXPERIMENTAL SETTING

The experiments were conducted in an office at ISI during normal working hours (see Figure 4-2). Subjects were brought into the experimental setting one at a time. The

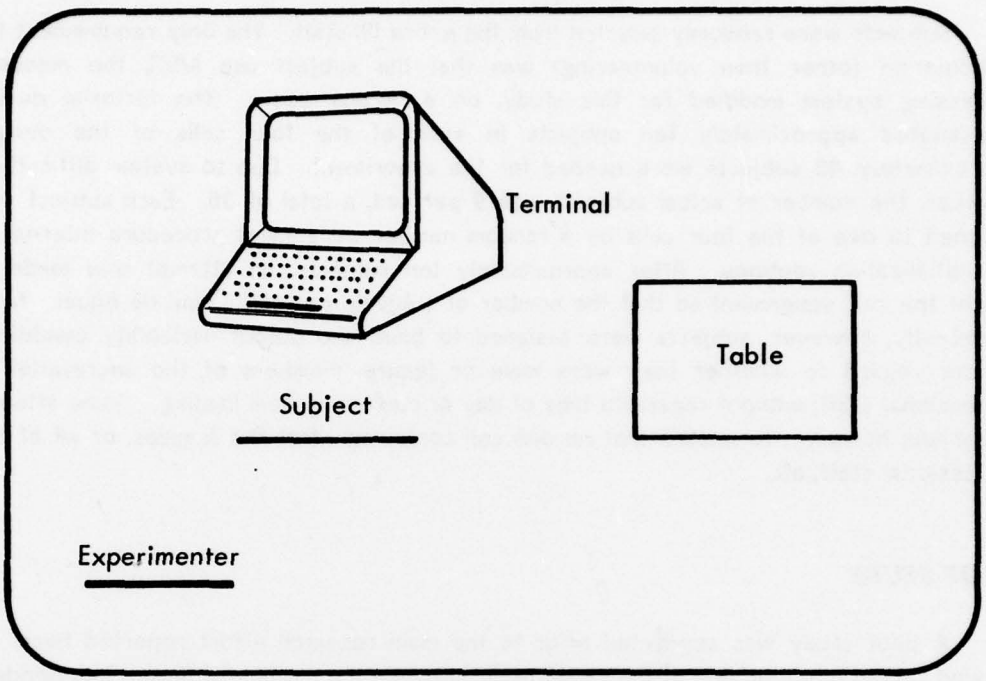


Figure 4-2. Experimental arrangement

room contained the usual ISI office furniture, including a Hewlett-Packard 2640A CRT and keyboard (HP), the computer terminal which all participants in the experiments ordinarily use for a large part of their daily activities, and a table with answer sheets for writing responses to the series of tasks to be performed. The HP includes a 24 line by 80 character (5 inch by 10 inch) rectangular CRT display and a separate keyboard attached to the display by a connecting cable. The normal display rate is switch-selectable from 110 baud to 2400 baud. At ISI, terminals are used at the 2400 baud rate (approximately 240 characters per second). To simulate the 1200 baud display rate used in these experiments, each line of output was interleaved with "null" characters, which produce no output on the screen, but have a bit string and are handled as an ordinary character, in order to increase the display time for a given output string by a factor of two.

As the subject entered the office, he or she was given a set of instructions on the use of the system as well as the answer sheets for writing the answers to the tasks. (See Appendix 1 for the instructional materials given the subjects and the instructions read them by the experimenter.) The subject was invited to sit down in front of the terminal and was read the instructions by the experimenter. After being advised of his or her right to leave at any time, the subject was given an additional brief description of the nature and intent of the research. Subjects were then told that the first two tasks were sample tasks, and that they could use their time working on them to become familiar with the new commands available in the Travel Message System that are not included in the

usual ISI version of MSG. They were also told that no data was being taken during the sample tasks, but that during later tasks, data would be taken on line, and that they were to work as quickly and efficiently as possible.

They used the time on the sample tasks to experiment with commands in the system which they might not ordinarily use in their routine, day-to-day message retrieval procedures. In particular, most do not use the Multiple search (boolean) requests on a regular basis; the sample tasks allowed them the opportunity to refresh their knowledge of the operation of the boolean search requests.

For each of the sample tasks, and each of the actual tasks, the task question appeared in a reserved area at the top of the screen, where it would remain until the subject pressed the "N" key to go to the next task. All requested output would then appear below the reserved area and would scroll in the normal manner. Figure 4-3 indicates the way the screen looked to the subject with the first sample task description in the reserved area, and sample output in the working area.

Each task required the subject to read or count a number of messages and write the appropriate information on the answer sheets provided. Because subjects had a good deal of familiarity with the system in their daily use, they required little help in the use of the system during the experimental session. The experimenter was available to provide

S1) HOW MANY REQUESTS WERE MADE TO THE TRAVEL DEPARTMENT for travel to San Diego? Follow the instructions on the answer sheet to answer this question.

****

<- headers destination or date string: san diego

8	Larry Miller	Travel, San Diego, April 2, a.m.
11	Jane Doe	Travel, San Diego, March 26 p.m.
24	Larry Miller	Travel, San Diego, March 25 a.m.
39	Alan Schwartz	Travel, San Diego, May 26 a.m.
46	T. Smith	Travel, San Diego, May 15 a.m.
51	John Wilson	Travel, San Diego, March 16 p.m.
55	Sim Farar	Travel, San Diego, March 13 p.m.
57	Alan Schwartz	Travel, San Diego, Feb.13 p.m.
69	Alan Schwartz	Travel, San Diego, Feb.26 a.m.
81	John Wilson	Travel, San Diego, March 20 p.m.
90	Bob Wilson	Travel, San Diego, June7 p.m.
92	Bob Wilson	Travel, San Diego, March 17 p.m.
97	David Simpson	Travel, San Diego, March 11 p.m.
101	Sim Farar	Travel, San Diego, April 2 a.m.
122	Sim Farar	Travel, San Diego, June 8 p.m.
130	David Simpson	Travel, San Diego, Jan. 21 a.m.
137	Larry Miller	Travel, San Diego, May 10 a.m.
155	Bob Wilson	Travel, San Diego, Feb. 11 a.m.

Figure 4-3. Travel Message System display after completing first sample task

help on the use of the functions if the subject requested it. Essentially, the experimenter simulated the help function normally available in MSG. The usual help facility, however, was not made available during these sessions because of programming peculiarities.

After the subject completed a task and was satisfied with his answer, he pressed the "N" key to go to the next task. After all of the tasks were completed, the instructions for the post-test questionnaire appeared on the screen. The subject completed the questionnaire, including an open-ended question which allowed him to express his general comments on the system and, in particular, to comment on any areas about which he might have felt strongly but which were not adequately covered by the previous questions. In particular, though it was not a part of the variables of this study, this final question offered an opportunity for subjects to express their ideas on ways to improve the system for the kinds of tasks performed in this study.

Each subject's total time in the experimental session varied with the particular combination of independent variables experienced; the average was about 1 to 1-1/2 hours.

A brief comment is in order at this point on the effect of a heavily loaded system upon the subjects. For a number of the subjects in the high variability versions of the system, the experiment was not entirely pleasant. Many of these subjects felt, and rightly so, that the system did not provide an adequate range of functions to efficiently work the tasks. These opinions are mirrored in the responses to a number of the questions in the post-test questionnaire; these are discussed in greater detail in the next section. However, it is useful to point out that it is in a heavily loaded system that the inadequacies become noticeable and burdensome. It appears that it is necessary to stress a system in a reasonable operating environment in order to better perceive its usefulness and its shortcomings. A further discussion of this point is made in Chapter 6.

DATA ANALYSIS

Subjects were randomly assigned to one of the four conditions of the factorial design, as described above. The theoretical basis for analyzing the relationship between these independent variables and the performance measurements of this study were described in the previous chapter. The factorial design, with repeated measures on each subject across the volume levels, lends itself to analysis by the classical analysis of variance methods. Analysis of variance provides a means of partitioning the total variance in the dependent measures into that which can be accounted for by differences in the independent variables (as well as that which cannot be accounted for by these differences). The ratio of the between groups variance to the within cell variance provides a test of the probability that the *observed* differences between the means of the cells in the factorial design are due to chance differences only.

The analysis of variance provides a means of testing for the main effects and the interactions as described above. Specifically, analysis of variance was used to test the effects of the independent variables upon the performance measures separately. Three separate analyses were required to test the effects on the time to complete the tasks, the CPU time used, and the number of functions used to complete the tasks. The data was analyzed with the performance measures totalled for the entire session, for low and high volume tasks.

The analysis of the data of these experiments involve the classical analysis of variance for differences in the main effects (main effects for output variability, output baud rate, output volume), and for the significance of the interaction effects. Classical analysis of variance does not provide the mechanism for testing the significance of the effects of a group of independent variables upon a group of dependent variables. Put another way, if the dependent or performance measure is viewed as an n-dimensional quantity, then multivariate techniques are necessary. The analyses carried out on the data of this study includes multidimensional analyses of the relationship between the independent variables and the n-dimensional performance measures; the results are detailed in the next section. Specific methods used were a multidimensional analysis of variance to assess the effect of output variability alone on the n-dimensional performance measures, and canonical correlation, to assess the relationship between the two sets of variables.

VALIDITY

Internal Validity

Campbell and Stanley [1963] and others provide a framework within which the necessary control on the experimental design may be exercised in order to better assure internal and external validity. Specifically, *internal validity* refers to whether observed changes or differences between groups may reasonably be ascribed to changes or differences in the independent variables of the study. The possiblity of uncontrolled, unanticipated, or unknown differences between the groups in the experimental design must be considered whenever an experimental design is analyzed. Following is a list of eight extraneous variables identified by Campbell and Stanley which, if not controlled for in the design, may produce effects which are confounded with the experimental variables.

1) HISTORY--***Specific events occurring between testing in addition to the experimental variables.***

Not applicable, since the design of this study did not involve a test, re-test situation.

2) MATURATION--***Processes within subjects operating as a function of the passage of time (fatigue, hunger, boredom, etc.).***

The experimental session was short enough and involved types of tasks sufficiently similar to what subjects perform in their usual work that fatigue and hunger effects are not reasonable. Boredom was a problem with some of the subjects in at least one of the experimental conditions (high variability coupled with 1200 baud output rate). However, this change in the subject's attitude, from interest to boredom, was considered to be of importance in evaluating the subject's overall response to the system and its environment. Thus, though boredom itself was not measured directly, its consequences -- as indicated by the answers to the post-test questionnaire -- were of interest.

3) TESTING--*The effects of testing upon subsequent testing.*

For those subjects who were involved with the pre-test, there was a sufficient amount of time between testing sessions (approximately four months) so that they approached the sessions with no *apparent* carry-over effects. In any event, subjects were assigned randomly to groups in the design, so that there appears not to have been any selection bias. See item 6 below.

4) INSTRUMENTATION--*Changes in obtained measurement due to changes in instrument calibration, or changes in the observers or judges.*

All data was taken on-line. Specifically, the timing measurements, CPU time used, and count of keystrokes was compiled internally by the program. Also, the post-test questionnaire was administered on-line: subjects typed their answers into the machine and the answers were written to a file by the program.

5) STATISTICAL REGRESSION--*Also known as Regression Towards the Mean. When groups are selected on the basis of extreme scores on a pre-test, we expect, regardless of any treatment differences, that each will re-test closer to the initial mean.*

The subjects were not selected on the basis of any pre-test. Each subject was randomly assigned to one of four experimental groups at the beginning of the session.

6) SELECTION--*Bias resulting from differential selection of subjects for the experimental groups.*

Random assignment of subjects to treatment groups is an effective way of avoiding the Selection bias. By having the computer assign the subject to an experimental group, experimenter bias in the selection was avoided. To further reduce the possibility of having all of the females in one group, or all of the secretaries, etc., some attempt was made to balance the distribution of identifiable subject categories amongst the four comparison groups: male, female, secretarial, professional.

7) EXPERIMENTAL MORTALITY--*Differential loss of subjects from the experimental groups.*

Two subjects were lost, in that their data was invalid due to system difficulties. This problem in internal validity, however, is of greater concern in a test, re-test situation. No data is included in the task data or the analysis of the post-test questionnaire of those subjects who were unable to complete the session. Subject mortality is of concern also, when the reason for subjects dropping out is loss of interest, low scores, etc.

8) SELECTION-MATURATION INTERACTION -- *Interaction effects between any of the above variables, which may be mistaken for the effects of the experimental variables.*

This confounding effect is again of greatest concern in test, re-test designs.

Other sources of error which may have an effect on the validity of the experimental design include experimenter bias, reactive measures effects and rating errors. Attempts were made to control for each of these. In particular, to control for differential effects of experimenter bias, a consistent set of instructions was read to each of the subjects. Though the experimenter could determine the experimental group to which the subject was assigned, he was careful to maintain a non-obtrusive attitude towards all subjects. Because the session was essentially self-paced, little or no experimenter intervention was required or necessary, except for those specific cases in which the subject requested help. In such cases, the experimenter attempted to simulate the normal help facility available in the message system.

Rating errors were not a factor in these experiments, since all data was taken on-line. A more serious source of error was the effects of the experimental environment per se on the subjects. In particular, the possibility of guinea pig effects -- whereby the measuring and testing process itself, with the subject in an environment involving an observer -- changes the respondent and biases the results. However, since subjects were assigned randomly to groups, and the potentiality of guinea pig effects applied equally to all subjects, one may conclude that there would need be a selection, environment interaction for the guinea pig effect to influence the comparisons between treatment groups.

The same problem exists for role selection effects, where a subject may assume a role different from his natural behavior in situations similar to, but outside, the research setting. Consistent instructions to subjects, including the statement that the tasks are not meant to be a test of individual performance, but rather of the performance of the total man machine interaction, help to minimize this source of error. There is no doubt, however, that the use of reactive measures, an environment with an observer and a known test situation, may bias the results of this research. To the extent that the groups experienced essentially identical experimental environments, except for the differences in

the independent variables, one may reduce the consequences of reactive measures. The desirability of controlled observation, similar system parameters, lack of distractions, etc., make the reactive measures design seem the appropriate one. It would be useful in future research to be able to take the measurements unobtrusively, perhaps by gathering timing and other data on-line in the daily activities of users of a particular programming system. Instrumented versions of interactive programs are not uncommon. The difficulty, however, is in controlling the working environment in order to make useful interpretations of the data gathered and to provide meaningful controls so that useful performance comparisons may be made.

External Validity

Factors influencing the external validity, or generalizability, of the research may be broken into two categories: those factors which relate to the population from which the subjects were sampled, and those which relate to the realm of interactive computer systems, of which the Travel Message System is one example. In fact, of course, the two concerns are similar. The subjects for this study were selected from the staff at ISI. At the lowest level, one may say that they are representative of the members of the ISI staff who have had experience with MSG. It is reasonable to conclude that there is nothing unique about the hiring practices of ISI, or the use of MSG, that would forbid a generalization of the subjects to a larger group. A better description of the population from which the subjects were drawn would be those who use interactive computer data base search and retrieval programs on a regular basis, where the interaction device is a 1200-2400 baud CRT and keyboard. Subjects who utilize slower TTY (down to 110 baud) devices as their usual means of interacting with a machine may have different expectations of the speed and variability of the computer output. As the cost of mechanical devices such as TTYs increase, and the cost of CRTs decrease, more and more interactive computer users will involved in program development, text editing, or other non-search or retrieval-oriented tasks may similarly bring a different set of expectations to the interactive environment.

5. RESULTS

Subjects were randomly assigned to one of four groups in the factorial design described in the previous section. As indicated below, these groups are 1200 baud, low output variability, 1200 baud, high output variability, and 2400 baud, low and high output variability (see Figure 5-1).

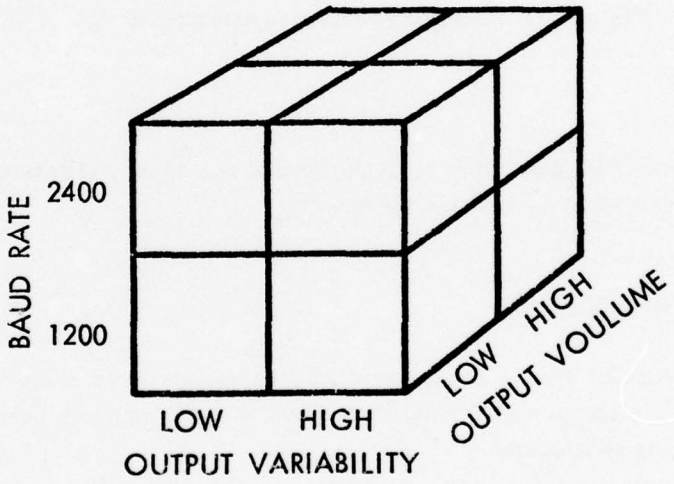


Figure 5-1. 2 x 2 x 2 factorial design

In addition to the two independent variables, baud rate and output variability, there was a third variable, output volume. Subjects were measured at more than one volume level. The final experimental design, as discussed earlier, is a 2x2x2 factorial design, with repeated measures on one of the factors (output volume).

One of the reasons for using the repeated measures design is to reduce the within-cell or error variation. Another reason to use this design is economy of subjects. Though there exists the possibility of confounding main effects with subject differences in the mixed design utilized here, random assignment of subjects to groups leads one to reject this possibility [Winer, 1971]. In the simplified design below (Figure 5-2), any observed differences in the criterion (dependent) variable across A could be ascribed to either the effect of A upon subjects or the consistent differences between subject group G1 and subject group G2.

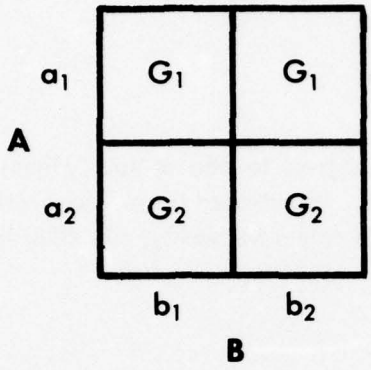


Figure 5-2. Simplified repeated measures design

Because of random selection of subjects, we will ascribe any differences across A to be A effects rather than subject difference effects.

ORGANIZATION OF SECTION

A brief discussion of the organization of this section is in order. First will be presented the results concerning the series of tasks that the subjects performed. These support the following conclusions:

CONCLUSION 1
There is a significant effect for output variability on user performance.

CONCLUSION 2
The effects of changes in the output rate on user performance are not significant.

CONCLUSION 3
There is a significant effect for output volume on user performance.

CONCLUSION 4
There is a differential effect for output variability on performance at different volume levels. At high output volume, the effect of increased variability is greater than at low volumes.

CONCLUSION 5
Those subjects experiencing the 1200 baud, low variability version of the system

performed significantly better than those experiencing the 2400 baud, high variability version.

These conclusions and supporting data are presented below.

Secondly will be presented the results of analysis of the post-test questionnaire. These results support the following conclusions:

CONCLUSION 1
Those subjects who were in the high variability conditions had a significantly poorer view of the interactive computer system and its environment than those in the low variability conditions.

CONCLUSION 2
There is a significant differential (interaction) effect between variability and output display rate upon the user's view and tolerance of the system.

CONCLUSION 3
There is a significant difference in the user's attitude towards the system and its environment between those who received the low variance, 1200 baud version, and the high variance, 2400 baud version.

CONCLUSION 4
There is no significant difference in the user's attitude towards the system and its environment between those in the 1200 baud version and those in the 2400 baud version.

CONCLUSION 5
Those subjects who felt that the system was too slow, or too variable in response, were also less satisfied with other (non-manipulated) features of the interactive environment.

TASK RESULTS

Data was pooled across all 11 tasks. Additionally, tasks were divided into ones requiring low output volume and those requiring high output volume. The task with the median output volume was eliminated in order to further enforce the high-low dichotomy.

The main conclusion, that there is a significant effect for output variability on user performance, is supported, $p<.05$, across all volume levels. Analysis of variance summary tables for the repeated measures design, with repeated measures across the two volume levels (Table 5-1), appear below, and support this conclusion.

TABLE 5-1(a)
ANALYSIS OF VARIANCE SUMMARY TABLE

(Independent Variable: Total time in seconds to complete tasks)

SOURCE	SS	df	MS	F	p
Variability	1370340	1	1370340	6.10	<.05
Baud	398	1	398	<1.00	N.S.
Var x Baud	54	1	54	<1.00	N.S.
Error (bet)	7196608	32	224894		
Vol	17743917	1	17743917	233.30	<.01
Var x Vol	583020	1	583020	7.67	<.01
Baud x Vol	3160	1	3160	<1.00	N.S.
Var x Vol x Baud	2962	1	3296	<1.00	N.S.
Error(w/in)	2434000	32	76063		

TABLE 5-1(b)
ANALYSIS OF VARIANCE SUMMARY TABLE

(Independent Variable: CPU time used)

SOURCE	SS	df	MS	F	p
Variability	9.90	1	9.90	<1.00	N.S.
Baud	1245.80	1	1245.80	19.60	<.01
Var x Baud	4.7	1	4.70	<1.00	N.S.
Error (bet)	2042.60	32	63.80		
Vol	1099.00	1	1099.00	17.90	<.01
Var x Vol	26.00	1	26.00	<1.00	N.S.
Baud x Vol	599.20	1	599.20	9.80	<.01
Var x Baud x Vol	3.30	1	3.30	<1.00	N.S.
Error(w/in)	1962.00	32	61.30		

TABLE 5-1(c)
ANALYSIS OF VARIANCE SUMMARY TABLE

(Independent Variable: Keystrokes used)

SOURCE	SS	df	MS	F	p
Variability	80.20	1	80.20	<1.00	N.S.
Baud	288.00	1	288.0	<1.00	N.S.
Var x Baud	312.50	1	312.50	<1.00	N.S.
Error (bet)	14095.60	32	440.5		
Vol	53.40	1	53.40	<1.00	N.S.
Var x Vol	450.00	1	450.00	1.50	N.S.
Baud x Vol	555.60	1	555.60	1.90	N.S.
Var x Baud x Vol	501.40	1	501.40	1.70	N.S.
Error(w/in)	9570.00	32	299.10		

In Table 5-1(a), note that the effects for Baud, Var x Baud, Baud x Vol, and the triple interaction are not significant ($p>.05$).

It is instructive at this point to view the data as a 2x2 factorial design, as indicated in Figure 5-3, combining low and high baud groups. The mean times (in seconds) to complete the tasks are indicated in their proper cells, and the results are plotted in Figure

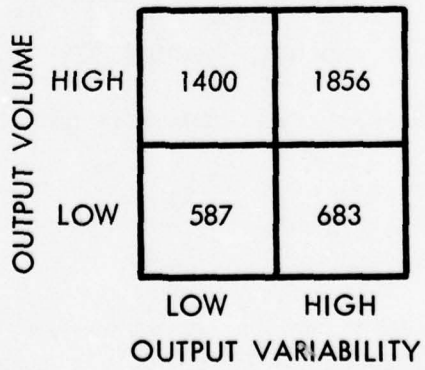


Figure 5-3. 2 x 2 design--1200 and 2400 baud combined

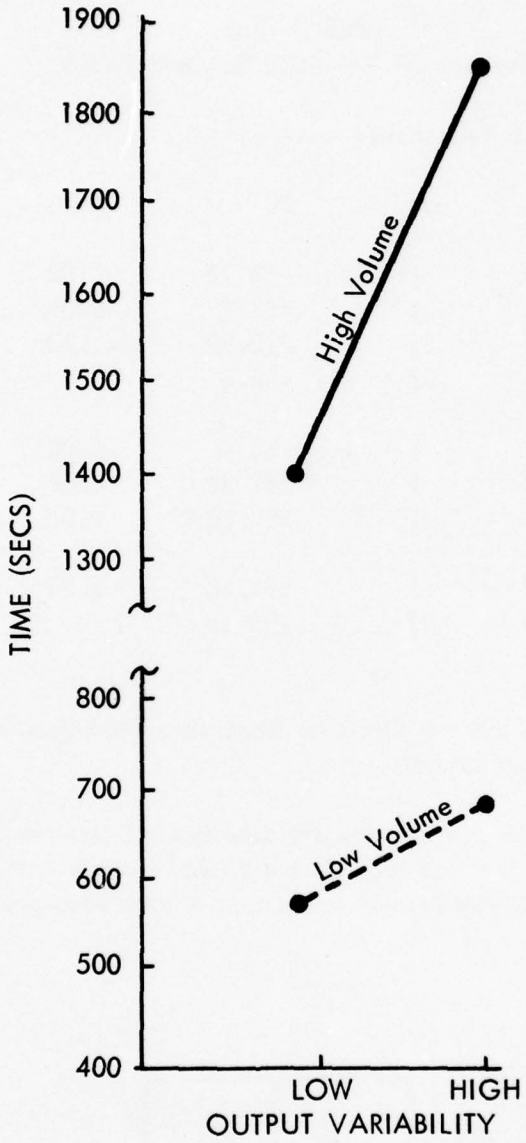


Figure 5-4. Plot of time vs. variability for two volume levels

5-4. This more clearly shows the effects of increased output variability and increased output volume.

The nominal baud rates for the low and high variability conditions and the low and high baud rate conditions are indicated in Figure 5-5. The numbers in brackets indicate the average baud rate over the row or column, as appropriate.

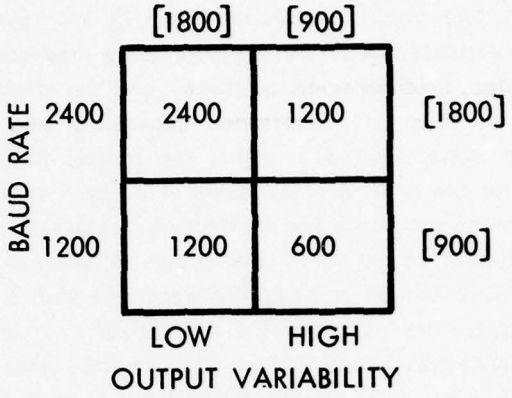


Figure 5-5. Nominal baud rates

Since the high variability conditions yielded a nominal baud rate of 1800, and the low variability condition one of 900, we might expect differences in performance to accompany these differences. Similarly, the average difference in baud rate is also 900 vs. 1800. But here, no significant performance differences are observed, $p>.05$. Thus the significant performance difference between the low and high output variability groups can *not* be accounted for by differences in nominal output rates. The implications of this result will be discussed later.

In Figure 5-5, note that there are two cells with identical nominal output rates--2400, high variability, and 1200, low variability. Each yields a nominal 1200 baud output rate. It seems clear that any performance differences between these two cells can be attributed to output variability differences only. Furthermore, comparing these two cells allows a test of which users would prefer: smooth but slow output, or jerky but fast. Additionally, the post-test questionnaire offered an opportunity to compare attitudes of users, as well as their behavior (performance). A t-test was performed comparing these two cells of the factorial design. The mean difference in time to complete the tasks between cells was significant: $t = 4.28$, $p<.01$.

Performance measures other than time to complete tasks were taken, including CPU time used and number of keystrokes or functions used to complete tasks. The analysis of variance summary tables for these variables are presented above in Tables 5-1(b) and 5-1(c). The effects of changes in the output baud rate and the output variability on CPU time used can be accounted for by the algorithms necessary to implement the 1200 baud version and the high variability version.

For each subject on each task, three performance measures were obtained: time to complete task i, T_i; CPU time used in completing task i, C_i; and keystrokes used, K_i. In the traditional analysis of variance, these three performance measures are considered to be independent, uncorrelated, unidimensional quantities, and the effects that the independent variables may have on them are determined separately and independently. Three separate analyses are done, each determining the effects of the set of independent variables on only one of the dependent variables at a time. When there is a reasonable interpretation which can be placed on the performance measures in a metric sense, then it becomes useful to attempt to interpret the observed differences. In the task data presented here, the interpretation is straightforward: the high output variability groups took more time to complete the tasks than the low variability groups. We can reasonably say that their performance was "better." If we examine the number of keystrokes used in performing the tasks, it is not clear whether using more keystrokes should be considered "better" performance or "worse." In this case, it would be necessary to attach meanings to the number of keystrokes used *before* the data analysis is accomplished. Post-hoc (or really, ad-hoc) reasoning is inappropriate. Since there was no strong motivation for stating a priori whether more, or fewer, keystrokes were "better," we leave the presentation of the result in the form of stating that there was, or was not, an observed difference. In the analysis of the post-test questionnaire, reported below, this same problem of interpretation also appears. However, there it is possible to make a priori statements concerning the meaning of response differences. This is reported in detail below.

Discussion and Analysis of Test Results

The variables of interest in this study were output baud rate and output rate variability. A third variable, output volume, was used to control for differences that large vs. small amounts of material to be read would have on the performance measures. The rationale for selecting these variables, and the performance measures, is discussed earlier. To the extent that these objective measurable quantities capture the effects upon the user of differences in system parameters, they can be described as good, or useful indicators of effectiveness of the interactive system from the user's point of view.

In this study there are two sets of performance measures: those taken during the tasks -- time, CPU, and functions -- designed to objectively measure the user's performance, and the attitude survey (post-test questionnaire), designed to measure the user's attitude towards the system and the interactive environment.

The effects of the independent variables on the performance measures will be discussed individually, then their combined effects will be discussed.

Main Effects
Variability Effects

There is a significant difference in time to complete the tasks, across the output variability. Further comparisons of the two conditions, 2400 baud/high variability vs. 1200 baud/low variability were done in order to ascertain the possibility of confounding effects. These two cells produced equivalent nominal output rates: 1200 baud (see Figure 5-5). As described earlier, the variability algorithm was designed such that the total time to display N characters on the screen would be approximately double the amount of time to display the same N characters without the variability. Therefore, any performance differences between these two conditions can reasonably be ascribed to variability of output differences rather than total output time differences. The result of the comparison of performance differences between these two conditions is that there is a significant difference in the time to complete the tasks. The amount of CPU time used also varies, but this may be attributed to the additional processing needed to implement the 1200 baud display rate, and the high output variability. There was no significant difference, however, in the number of keystrokes used in performing the tasks.

The conclusion is that increasing the variability of computer output significantly decreased user performance in the interactive tasks. To the extent that the test population for these experiments represents a broader category of potential interactive systems users, and the test system is representative of a broader class of interactive systems, we may conclude that variability in display rate has per se a detrimental effect on user performance. The question of the generalizability of the results of this study, the issues of reliability and internal and external validity, were introduced in the previous chapter and will be discussed further in the next.

Baud Rate Effects

In contrast to the significant effects on performance of changes in output variability, presenting the requested information in the interactive travel message system at 1200 baud vs. 2400 baud produced *no* significant differences in performance. In particular, even though the nominal baud rate was 1800 for the 2400 baud groups and 900 for the 1200 baud groups (see Figure 5-5), there was *no* significant difference in times to complete the tasks for the two groups. This result holds across high volume as well as low volume tasks (i.e., the baud x volume interaction was not significant). It appears that both 1200 and 2400 baud display rates are faster than the typical subject can read, so that time to read a page of material depends on the individual's reading speed rather than system display rates. One would then conjecture a plateau in the curve of time to read a screen of text vs. display rate, somewhat like the one in Figure 5-6. Apparently the plateau is reached at display rates of less than 1200 baud. In fact, 1200 baud corresponds to approximately 1200 words per minute, a rate considerably faster than the rate at which the average person reads.

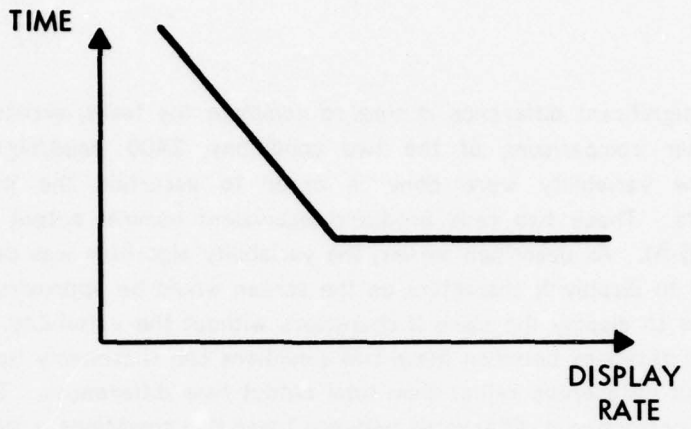


Figure 5-6. Graph of time to read screenful of material vs. display rate

However, the result is still somewhat curious. A number of tasks required the subject to visually search message bodies for particular names. While it would seem reasonable to expect that doubling the display rate should lead to shorter times in completing those kinds of tasks, this was not observed to occur. We may conclude that doubling the display rate from 1200 to 2400 baud does not produce improved performance for the subjects and system of these experiments. It should also be noted that the total time to present the typical amount of material was only about seven percent of the total time subjects needed to perform the individual tasks. For example, for the high volume tasks, average total output was approximately 2000 characters. At 1200 baud, this would require approximately 20 seconds to display this amount of material on the screen. However, from Figure 5-4, it is observed that the time to complete the high output volume tasks varied from 1400 to 1900 seconds, about 280 to 380 seconds for each of the five tasks, depending on output variability. Similar results hold for the low output volume tasks.

This result makes the one involving output variability seem that much stronger. Examining Figure 5-5 again, we note that there was a difference in average display rates across the two variability conditions. It is not immediately clear whether the effect for variability might be confounded with the average display rate. That there is no significant difference across baud rate leads us to reject that possibility.

Interaction Effects

Of the four interaction terms in the analysis of variance (variability x baud, variability x volume, baud x volume, and the triple interaction), only the variability x

volume effect is significant. The influence that increased output variability has on performance is greater at higher output volumes than at lower volumes. This result is not unexpected, since for this study the low output volume tasks required less than a screenful of text while the high volume tasks required considerably more. Apparently subjects could tolerate the higher variability in output rate if it occurred over a relatively short period of time. Also, in the high output volume tasks, a greater percentage of the task time was spent in reading output as opposed to the other actions required in the tasks. Though there has been little reported in the literature on the effects on reading speed or visual search of continuously varying the display rate, reaction time experiments tend to support the view that increasing the variability of the stimulus significantly increases response time (cf. Mackworth, 1970; Mostofsky, 1970; Davies and Tune, 1969).

POST-TEST QUESTIONNAIRE

A questionnaire (see Appendix 3) was administered to each subject on-line immediately following his or her series of tasks. Because the questionnaire was presented using the same values of the output parameters as for the tasks, subjects received a consistent view of the system. The questionnaire consisted of 18 questions relating to the complete system -- its speed, use of keyboard, display features, screen size, etc. -- and asked the subject to numerically rate the particular area of the question on a 5-point scale. The rating scale was presented to the subject before he or she began answering the questions. The following paragraph was presented:

> Please answer the following questions with a numerical rating in the range of 1-5. 1 = Very Poor, Unacceptable, etc. 5 = Excellent, Completely Acceptable, Easy to Use, etc. (1-2 implies a generally negative response, 4-5 a generally positive one.) However, a specific numerical scale will be given for each question.

Even though specific meanings were attached to the ratings for each question, all obeyed the same ordering: 1-2 indicated a negative reaction to the point or feature of the question, 3 a generally neutral response, and 4-5 a generally positive response. The questionnaire and scale selection were designed to follow standard practices in questionnaire and survey research instruments [Babbie, 1973]. The responses make at least an ordinal scale. Subjects were instructed to think of the response scale as representing a continuum and to answer anywhere in the range 1-5. Though not a ratio scale, and probably not an interval scale, a number of analytical techniques are available for analyzing the data. The analysis of the questionnaire data is presented later in this section.

The Questions

Not all of the questions were of immediate relevance to the tasks the subject performed, or to the variables of interest in this study. They were included, however, in order to provide a more complete view of the subject's overall impression of, and response to, the system as he or she had just experienced it.

The questions can be conceptually broken into groupings which correspond to different features of the interactive system. For example, questions 1-4 dealt with the use and the completeness of the system's commands. Questions 5-8 dealt with the physical aspects of the display: screen size, character size and shape, sufficiency and readability of output, etc. Questions 9-13 dealt with computer and printing speed, and variation in those speeds. Finally, 14-18 dealt with the overall utility of the system.

The following table (Table 5-2) presents the results for each question individually. The actual F ratios and probabilities are presented. Those which were significant beyond the 1 percent level ($p<.01$) are indicated by **. Those that were significant beyond the 5

TABLE 5-2
Significance of Individual Questions
in the Post-Test Questionnaire

Question	F_v	F_b	F_{vxb}	P_v	P_b	P_{vxb}
1	11.7	1.10	<1	**		
2	<1	<1	<1			
3	<1	<1	2.34			
4	6.38	6.38	<1	*	*	
5	1.92	1.44	<1			
6	8.16	1.44	<1	**		
7	<1	2.97	4.77			*
8	<1	<1	<1			
9	6.56	<1	1.81	*		
10	2.47	1.36	<1			
11	16.35	<1	<1	**		
12	17.31	<1	<1	**		
13	4.02	<1	<1	*		
14	4.93	<1	4.49	*		*
15	6.49	6.20	3.08	*	*	
16	1.78	3.94	<1			
17	1.16	2.26	<1			
18	7.96	<1	5.62	**		*

All with df = (1,26)

percent level ($p<.05$) are indicated by *. If in fact the data were uncorrelated, independent, we could make a probability statement of achieving N significant F ratios in M measures, at significance level p. Since our data does contain correlations, with the same subject providing answers on all questions, our probability statement is weakened.

The 18 questions making up the post-test questionnaire may be thought of as comprising an "index of satisfaction" of the user with the system. If we simply add up the responses of each subject, this sum may be considered the satisfaction index. Figure 5-7 presents the average response for the subjects in each of the four cells of the factorial design, while Figure 5-8 presents the results graphically. An analysis of variance was performed using this index as the dependent measure.

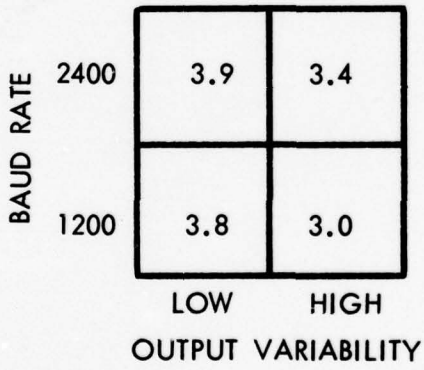


Figure 5-7. Mean responses for the 2 x 2 factorial design

The results of the analysis of variance are presented in the following table (Table 5-3(a)). We note the strong effect for output variability on the average answer to the post-test questionnaire.

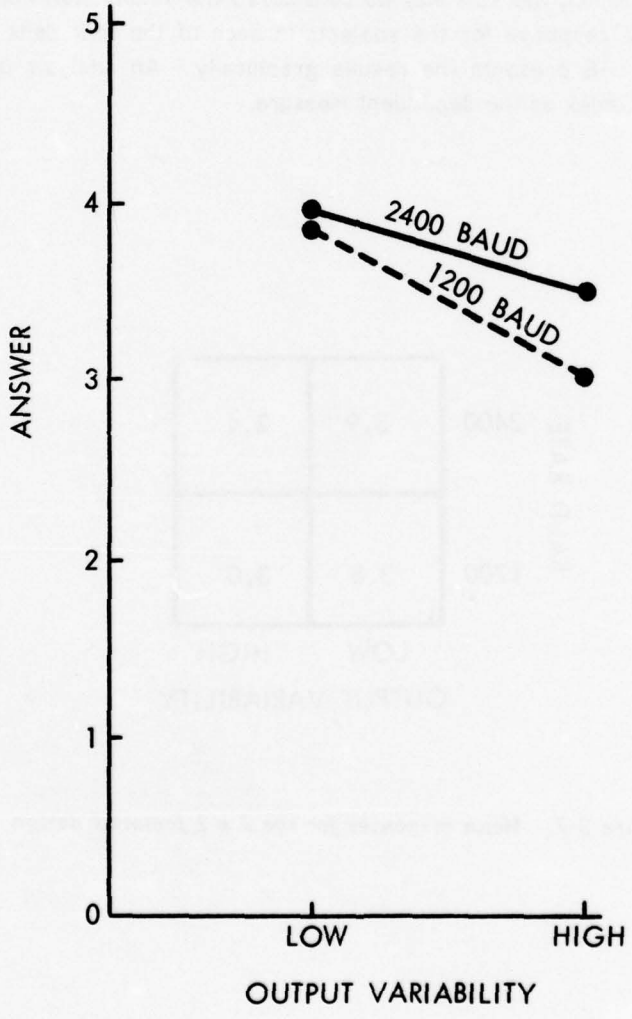


Figure 5-8. Graph of average response to post-test questionnaire vs. output variability for 1200 baud and 2400 baud

TABLE 5-3(a)
ANALYSIS OF VARIANCE SUMMARY TABLE

(Independent Variable: Average answer to 18 questions of post-test questionnaire)

SOURCE	SS	df	MS	F	p
Variability	2.77	1	2.77	19.4	<.01
Baud	0.43	1	0.43	3.0	N.S.
Var x Baud	0.29	1	0.29	2.1	N.S.
Error	3.40	24	0.14		

If we view the data as an 18-dimensional quantity then it is reasonable to compare not the sum (or average) of the answers, but rather the norm of the answer vector in 18-space. This norm is given by the square root of the sum of the square of the answers to the individual questions, $|Q| = (\Sigma\ q_i^2)^{1/2}$, where q_i is the answer to question i, |Q| is the norm of the subject's answers. *The analysis of variance summary table is presented* below (Table 5-3(b)), using |Q| as the dependent variable, and nominal output rate (baud) *and output rate variability as the independent variables.*

TABLE 5-3(b)
ANALYSIS OF VARIANCE SUMMARY TABLE

(Independent Variable: Norm of answers to 18 questions of post-test questionnaire)

SOURCE	SS	df	MS	F	p
Variability	0.14	1	0.14	4.67	<.05
Baud	0.12	1	0.12	4.00	N.S.
Var x Baud	0.08	1	0.08	2.67	N.S.
Error	0.76	24	0.03		

In both of the analyses, using either the sum of answers to the 18 questions or the norm of the answers, there is a significant effect for output rate variability but not for nominal rate differences. Both analyses allow an unequivocal view that users

experiencing the high variability versions of the system expressed a significantly lower view of the system, its commands, its display, its speed and its overall utility than those experiencing the low variability versions of the system. The questionnaire is further broken down later by questions relating to commands, display, speed and utility.

Interaction Effects

In questions 3, 7, 9, 14, 15, and 18 an interesting pattern of interaction effects was observed. Analyzing the questions individually does not provide adequate sensitivity as to whether this pattern represents random (uncontrolled error) effects in the data or can reasonably be ascribed to differences in the display variables of this study.

It is instructive to examine the nature of the V x B interaction. In *all* cases, the relationship between output baud rate (B), output variability (VAR), and group mean answer is as presented in Figure 5-9.

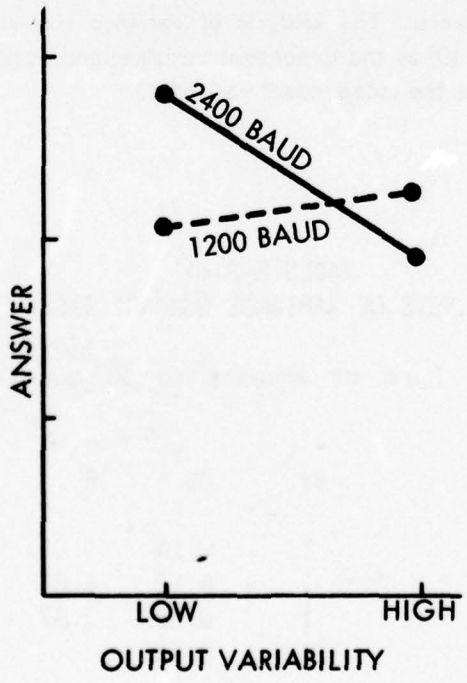


Figure 5-9. Graph of typical response to questions where interaction was significant vs. output variability

In each case, for the low variability condition, the subjects experiencing the 2400 baud version of the system had higher mean responses than those experiencing the 1200 baud version. This result is expected, and corresponds to the performance of these two groups in the tasks.

It is also noted in Figure 5-9 that the mean answers for the 2400 baud, high variability group is lower than that for the 2400 baud, low variability group. This result is also consistent with the performance data on the tasks. The interesting case, however, is that for those who experienced the 1200 baud version of the system. Those who experienced the 1200 baud, high variability version of the system had a pattern of *higher* mean responses than those experiencing the 1200 baud, low variability version of the system, or at least their answers were not significantly lower.

It is useful to examine the subjects' answers to individual questions within groups, and to the average response of the questions in each of the groups. For ease of later reference, the groupings of questions will be denoted [C] for 1-4 (commands), [D] for 5-8 (display), [S] for 9-13, (speed), and [U] for 14-18 (utility). Table 5-4 presents the analysis of variance summary tables for the average response within groups.

The analysis of variance summary tables for the individual questions have been presented previously (see Table 5-2). As noted earlier, if we examine the results for a main effect of output variability upon the answers, we find that there is a significant effect on questions 1, 4, 6, 9, 11-15, and 18.

When the answers are viewed as groups, each representing a common component of the interactive environment, we find that there are significant differences between subjects in low vs. high variability groups in the average response within the three groups [C], [S] and [U], but not for [D].

Figure 5-10 presents graphical results of the average response to the questions within the four groups ([C], [D], [S] and [U]), vs. output variability, for 1200 and 2400 baud.

If we (cautiously) embed our answers in a metric space, we are able to get a better insight into the nature of the differences within each group of questions. We may thus say that for those questions relating to system commands, [C], those experiencing the low output variability condition had an average response which was higher, and thus indicated better facility with the system, than those experiencing the high variability conditions. The same was true for those questions relating to system speed, [S], and overall system utility [U].

TABLE 5-4
ANALYSIS OF VARIANCE SUMMARY

SOURCE*	SS	df	MS	F	p
[C]					
Variability	1.38	1	1.38	3.93	~.05
Baud	0.14	1	0.14	0.40	N.S.
Var x Baud	0.39	1	0.39	1.10	N.S.
Error	8.45	24	0.35		
[D]					
Variability	0.93	1	0.93	2.45	N.S.
Baud	2.04	1	2.04	5.41	<.05
Var x Baud	0.13	1	0.13	0.35	N.S.
Error	9.06	24	0.38		
[S]					
Variability	7.26	1	7.26	16.24	<.01
Baud	0.04	1	0.04	0.09	N.S.
Var x Baud	0.04	1	0.04	0.09	N.S.
Error	10.7	24	0.45		
[U]					
Variability	2.64	1	2.64	13.63	<.01
Baud	0.83	1	0.83	4.31	<.05
Var x Baud	1.02	1	1.02	5.26	<.05
Error	4.65	24	0.19		

*Analysis of Variance summary tables for the Post-Test Questionnaire by groups: [C] -- questions 1-3; [D] -- questions 4-8; [S] -- questions 9-13; [U] -- questions 14-18

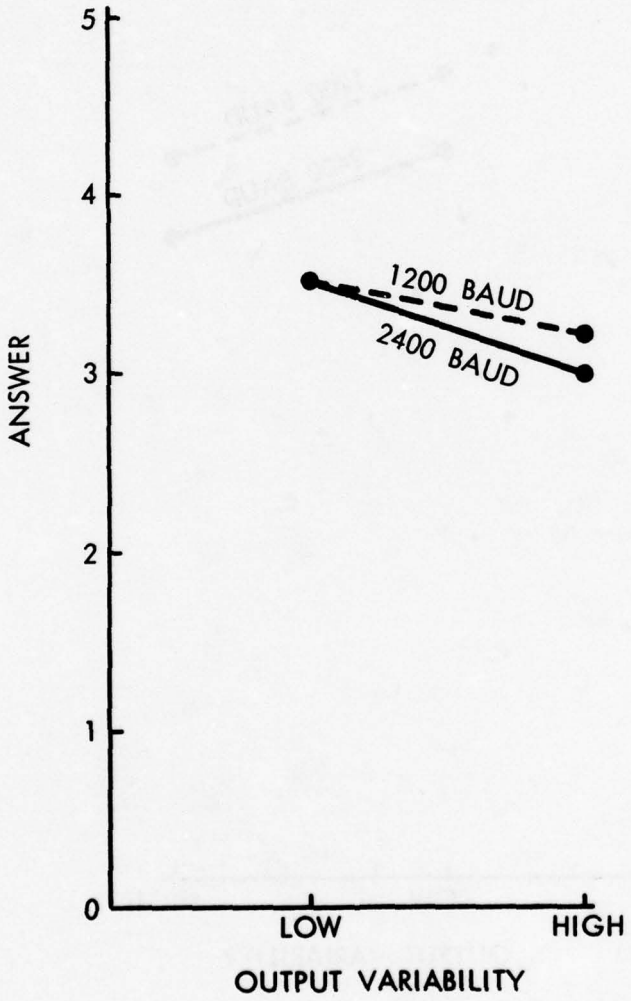


Figure 5-10(a). Graph of average response to the [C] questions vs. output variability

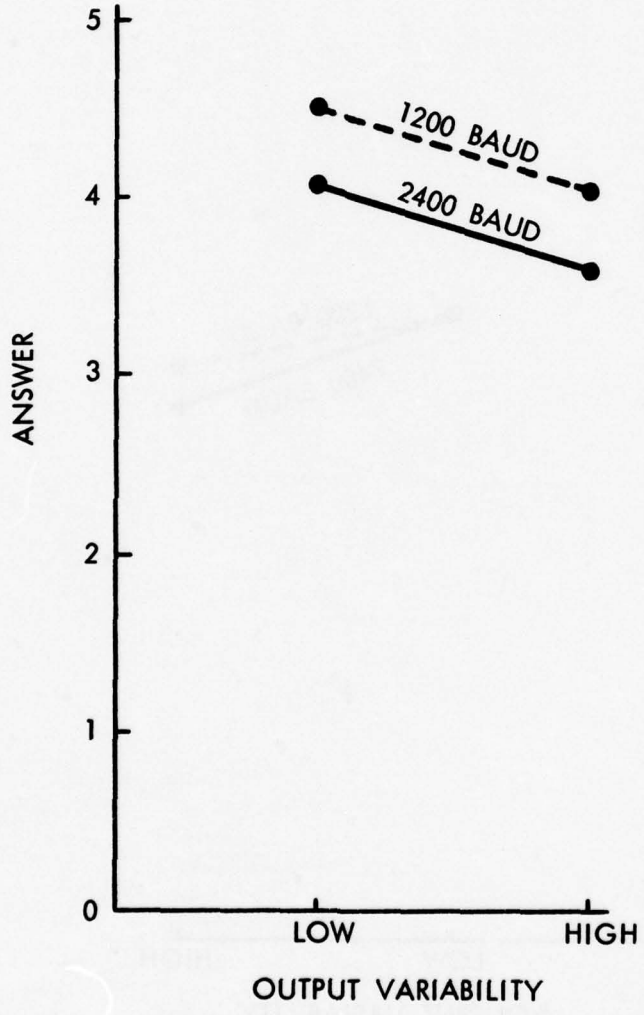


Figure 5-10(b). Graph of average response to the [D] questions vs. output variability

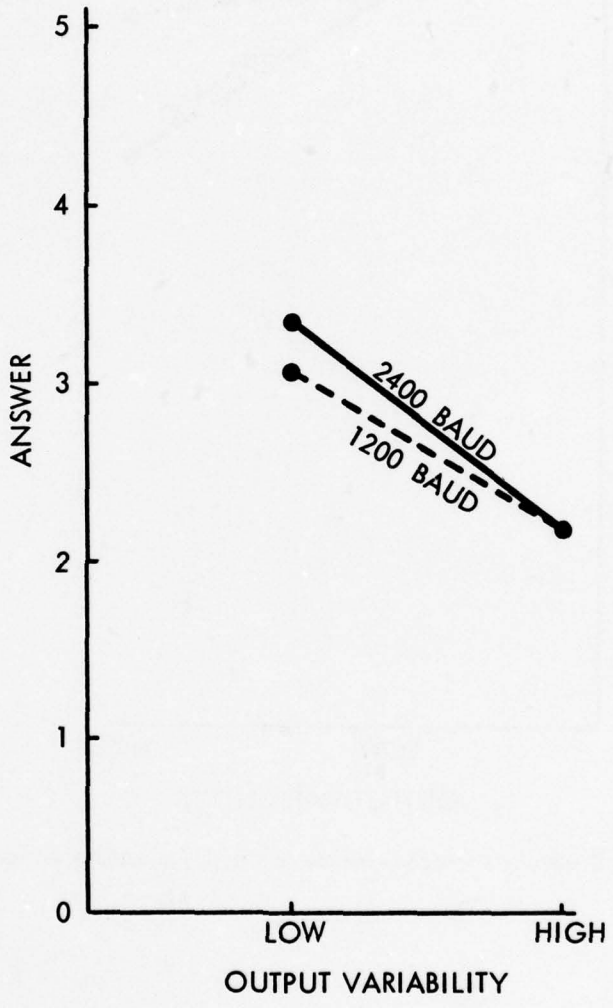


Figure 5-10(c). Graph of average response to the [S] questions vs. output variability

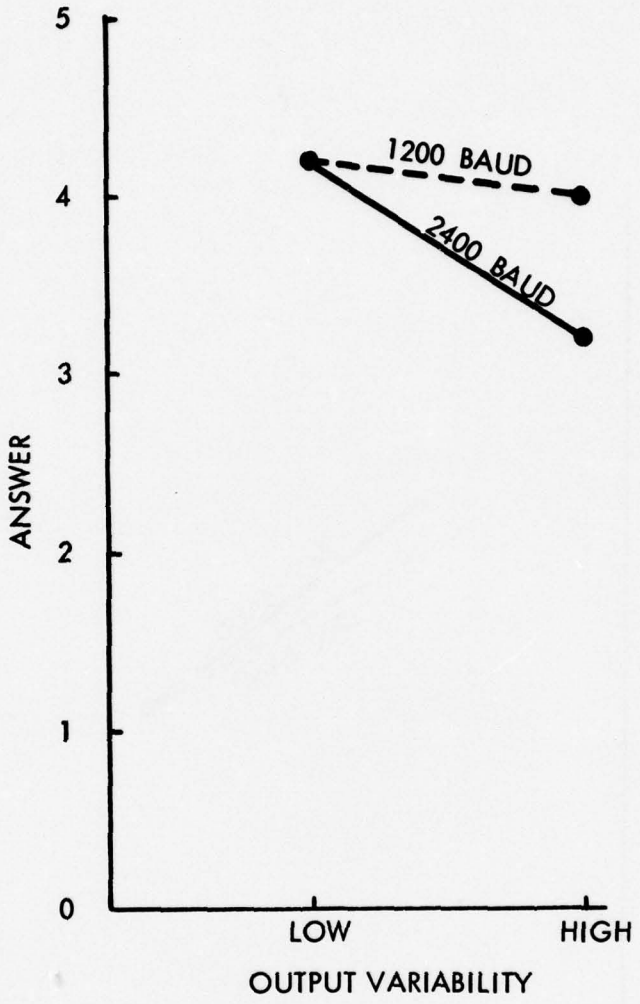


Figure 5-10(d). Graph of average response to the [U] questions vs. output variability

Discussion and Analysis of Post-Test Questionnaire

The above analyses of the subject's answers on the post-test questionnaire were made in order to provide insight into the the overall view of the system that the subject would have upon completing a typical search and display session with an interactive system. Some of the results need be examined in depth, since they may be at variance with intuition.

A fundamental conclusion, which is supported both by the task data and the questionnaire data, is that the nominal output baud rate, at 1200 baud vs. 2400 baud, has at best a very weak effect upon the user's performance and attitude towards the system. Specifically, over a number of tasks, involving both low and high output volumes, there was *no significant* performance difference between those receiving the 1200 baud version and those receiving the 2400 baud version. This result is observed across the low volume tasks, across the high volume tasks, and across the low output variability groups and the high output variability groups. On questions 3, 7, 9, 14, 15 and 18, an interaction as indicated in Figure 5-9 was obtained. In these questions, the average response of the 2400 baud group, in the low variability condition, was higher than in the 1200 baud group. In question 7, the 1200 baud group had a higher average response than the 2400 baud group, in the high variability condition. These six questions, 3, 7, 9, 14, 15 and 18, provide an interpretation problem. The result for the low variability condition is expected, and in virtually all of the questions, the 2400 baud answers were higher than the 1200 baud.

For those experiencing the high variability version of the system, the answers to the questionnaire were more interesting. In *all* of the questions except 10 (variation in computer system speed) and 12 (variation in printing speed), those experiencing the 1200 baud version of the system have *higher* answers than those experiencing the 2400 baud version. As indicated previously, these differences are generally not significant on an individual question basis, but when the questions are viewed in a multidimensional sense, these differences do tend to become significant.

The answers to the questions might tend to imply that if we are faced with a system which suffers from a great deal of variability in output rate (such as a heavily loaded interactive system, or a user interacting with a host computer through a communications interface processor), the nominal output rate should be adjusted to be about 1200 baud (if the selection is 1200 vs. 2400). This conclusion is not supported by a more thorough analysis of the interactive environment. Virtually all subjects in the high variability/1200 baud version voiced their frustration at the slow, unsteady nature of the output. One subject, in fact, refused to continue the experiment and another became openly hostile towards the experimenter. Their data are not included in the above analyses.

While subjects in the other conditions were able to maintain interest in the tasks and found the session a reasonable approximation to potential real-world tasks, those in the 1200 baud/high variability group had greater difficulty maintaining interest or motivation. Many felt that the tasks were putting undue burdens upon them, given the system that they had to accomplish the tasks. These conclusions seem reasonable because of two supporting results. Firstly, the experimenter notes of the comments of subjects during the sessions demonstrate the need for greater computer power (functions, speed) for this group (see Chapter 6 for further elaboration). Secondly, on some of the questions dealing with system functions and speed, subjects in this group had lower answers. Thus there appears to be an indication of deteriorating attitude towards the system by subjects in this group, which matches their decreased performance in the tasks.

Because subjects in this 1200 baud/high variability group had lower interest or motivation towards the tasks, because their frustration and dislike of the system appeared to increase during the experimental session, their general attitude towards the post-test questionnaire was one of relief that the agonizingly slow series of tasks had been completed. Their responses to questions were given not so much as a result of thoughtful consideration of the meaning of the question and the best selection of possible alternative answers, but rather as a result of a desire to hurry onto the next question and end the session as quickly as possible. It is clear from reading the questions that some do not make immediate contact with the exasperating nature of the system, and so require the subject to think about the answer and its implications. Questions such as 5 (screen large enough) or 16 (need for more materials on functions available) fit into this category. When the questions dealing with speed and variation (9-13) are examined, we find the expected relationship of this cell to the others. The table below (Table 5-5) presents the rank order of the 1200 baud, low variability group on the [S] questions, where 1 means this group had the highest mean response, 4 the lowest.

TABLE 5-5
Ranks of Questions in the Speed Group

Question	Rank of 1200 baud, low variability group
9	3
10	4
11	4
12	4
13	4

On those questions which most closely matched the strong feeling that the subjects in this group had towards the system, their answers reflect their behavior and their expressions during the session.

Examining the questions by groups, we notice a strong effect of output variability upon the attitudes of the users towards the system on three of the four question groups. Subjects experiencing the high variability conditions had a lower response index to the questions in the [C], [S] and [U] groups. Examining more closely the meaning of the [C], [D], [S] and [U] indices, the following are concluded:

(1) High output variability subjects perceived the command structure as less adequate to their needs. As indicated earlier, the experimental design assumed a fixed terminal type with its own display characteristics. The HP terminal used in these studies works on a scrolling method where each new line of output is presented on the bottom of the screen and all lines above scroll up. The topmost line is lost as each new line is appended at the bottom. Some of the apparent dissatisfaction with the command language (and the terminal itself) may be associated with the scrolling typical of the HP and other terminals.

(2) High output variability subjects were less satisfied with the physical display. Though this result is somewhat ambiguous, the general conclusion is that increasing the variability of the output display rate reduces the user's overall image of the system. Put another way, users with low variability of output seemed more likely to find the particular display satisfactory.

(3) Users who experienced the high output variability version were bothered by the slowness of the system, and noticed the reduced speed and the increased variability of the output rate. However, merely cutting the output rate in half (from 2400 baud to 1200 baud) did not produce a noticeable reduction in the answers for users in either the high or low variability groups.

(4) In general, subjects experiencing the high variability versions of the system had a lower view of the overall utility of the system, as evidenced by their average response to questions concerning the usefulness of the system, the desirability of using the system vs. performing the tasks by hand, need for more materials, and their overall rating of input to the computer and output from the computer.

Examining the intercorrelations between the answers to the questionnaire provides a useful insight into those parts of the system which are viewed as a whole. For example, looking at those questions which correlate significantly with the [S] questions, 9-13 (Table 5-6) allows one to identify those aspects of the interactive system with which a user is least satisfied as the system becomes more stressful. It would be expected, of course, that there would be significant correlations between questions within the [S] grouping, and this is observed. Identifying those questions outside of the [S] group which correlate with questions within the group, the following are concluded:

(1) Those who perceived the system as being slower had a significantly poorer view of the ease of using the commands and the overall utility of the system, felt a need for more materials on the system, and generally had a lower overall view of system output, than those who perceived the system as being relatively faster.

(2) Those who perceived the system as being relatively high in variability of output and processing speed had a significantly lower view of the ease of using the commands, found the brightness and size of the display screen less satisfactory, and found that the data presented was less sufficient for their needs than those who perceived the system as having relatively little variability in output or processing speed.

TABLE 5-6

Question	Questions with which it correlates significantly ($p<.05$)
9	10,11,12,13
10	4,5,8,9,12
11	9,12,13,14,18
12	1,5,9,10,13
13	9,11,12,16,18

6. CONCLUSIONS AND RECOMMENDATIONS

The major emphasis of this research is that there are a number of parameters of the man-machine interaction which affect the performance of the user. Specifically, it was hypothesized that changes in the nominal display rate of the presentation of computer output, and the variability in the display rate, would have significant effects on the performance and attitudes of the users of the man-machine system.

Reviewing the extensive statistical material presented in Chapter 5, we find that, contrary to initial conjecture, doubling the display rate from 1200 baud to 2400 baud has *no* apparent effect on user performance in the interactive tasks of this research. Also of interest, doubling the display rate from 1200 to 2400 baud has *no* apparent effect on the attitudes of the users towards the interactive system, its command structure, its display features (screen, characters, etc.), the speed of the system, or the overall utility of the system. In fact, in reviewing Table 5-2 again, we note that in only two of the questions in the post-test questionnaire (question 4 -- screen brightness, question 15 -- perform tasks by hand) was there a significant effect for baud rate on the subjects' answers. Even for the low output variability groups, the differences in the answers to the post-test questionnaire between those subjects receiving the 1200 baud version and those receiving the 2400 baud version were not significant.

When the effects of variability of display rate upon performance and attitude are evaluated, the situation is entirely different. In both the performance of the subjects on the interactive tasks and the attitude measures towards the system evaluated via the post-test questionnaire, subjects receiving high variability versions of the system performed significantly more poorly, and had a significantly poorer attitude towards the system, than those receiving the low variability versions. This effect is significant at both the 1200 baud and 2400 baud levels, and increases as the volume of output material increases. In fact, because of the significant interaction effect between output volume and output rate variability, we see that the effects upon performance are magnified at the greater output volumes. In the direct comparison between the two cells of the factorial design where the nominal output rates are the same (1200 baud low output variability vs. 2400 baud high output variability), the performance differences between the two groups is significant and quite strong at the high output volume level.

This research has found that doubling the display rate of system output to the user of an interactive message processing system does not improve performance, nor does it lead to an improved view of the system or attitude towards the system on the part of the user. What then are the effects of increasing system output, and what ways might be effective in both improving user performance and improving attitude? At this point, a

confounding problem occurs. It has been a (seemingly not unreasonable) assumption on the part of system designers that increasing the display rates leads to better performance in interactive systems. If we could guarantee that the variability in the display rate were held constant as the display rate were increased, the results of these experiments allow us to conclude that performance and attitude are not diminished. They may even be improved, though this does not appear to be the case in this research. So there is certainly no immediately apparent drawback to providing faster displays. However, as systems become heavily loaded, increased display rates are associated with increases in the variability. The actual display rate may not be improved, and the results of this research strongly demonstrate that performance is decreased and that user attitudes towards the system deteriorate. These conclusions are so strongly supported by the data presented (see Chapter 5) that a general recommendation to system designers would have to be that increasing output display rates should not be attempted without a corresponding increase in CPU power.

Subjects who received the high variability versions of the system experienced frustration and demonstrated a poorer view of the system and its environment in ways other than poorer performance and more negative questionnaire responses. For many of the subjects who participated in the experiments, their experiences with the system subjected them to conditions which were beyond their normal use of an interactive system. In particular, those subjects in the high variability conditions tended to express the opinion that they were utilizing a system through a heavily loaded terminal interface processor (TIP) port [Bolt Beranek and Newman, 1974]. Such was the agonizingly slow response of the system at times that the functions provided to perform the series of tasks were clearly inadequate. Those subjects receiving the low variability versions of the system did not tend to express such negative opinions of the system and generally found the experience to be a reasonable approximation of the types of tasks that they (or at least the person they were simulating) might have to perform on a regular basis. It is interesting to compare some of the responses to the open-ended question on the post-test questionnaire (see Appendix 3). The material presented below is not complete, but has been selected from both low and high output variability groups in order to illustrate the point.

Selected responses from subjects in the high output variability groups:

> Multiple mess. very useful but slightly confusing. Losing the page length in the middle of a search was a real pain. . .Output delays would become more of a pain as the proficiency increased. . .System was generally quite easy to use. Since questions were oriented toward info imbedded in the mess text it would be very useful to have the mess info either cross referenced or available to the command structure.

Another subject:

> I felt that the system was lacking in two basic areas. First of all, there was a deficiency in the functionality of the system. Too much data had to be scanned in order to answer the questions at hand. The addition of a few more commands would have greatly eased the tasks.
>
> Secondly, and far more important, was the fact that the system response was so incredibly poor. I felt myself thinking of the wasted time involved. It wasn't even consoling to think of the time needed to complete the tasks by hand. The thing that stuck out most to me was that I felt that I wasn't being useful (to myself or anyone else) when I was waiting on the machine. In that sense, poring over a listing by hand, and possibly taking longer to complete the task, would have been more satisfactory.

Another subject:

> The system (message system) would have been very easy to get used to -- definitions of commands etc. Would very much have liked a means to search the body of the messages for strings. Found the long delays from the computer made the job really dull -- it would have been more interesting if the information could have been gathered quickly and easily (the text string search).

Selected responses from subjects in the low output variability groups:

> The subcommand "Multiple" I feel should take more than a one-line string. For example I would like to be able to type From (string) and From (string) again and have the message service give me both From strings.
>
> I feel in general that this message program does its tasks well in that it gives the user the facts he needs quickly and reliably.

Another subject:

> I would like to see more commands similar to XED "Find" and "Search" or a method of reading all selected messages into a buffer to search for key words or phrases.

Another subject:

> There should have been more effort in being able to find names in the body of the message. Also, since there seemed to be no distinction between the sender and the rest of the people on the trip, they should have equal status for searching etc. It should also process dates into the standard format before the search (I don't want to remember the exact format -- especially which months are abrev. and which aren't).

> The system speed was acceptable. I just want more power.
>
> The output format should have been less verbose, i.e., find the people's names in the messages and just print those.

Quite often production systems are developed and implemented in an ad-hoc manner without adequate regard for the stresses that a heavily loaded system can place upon the performance and attitude of the user of the system. Apparently, from the nature of the responses to the open-ended question, a system designer can gain valuable insight into the needs, requirements, typical modes of interaction, etc., of users by observing a large number of potential users actually using the system or a simulated version of the system, under a number of different conditions. This conclusion was one of the motivations for the research reported in Heafner and Miller [1976]. In this work, the authors demonstrated the utility of observing and questioning a large number of potential users of a military message processing system. Unlike the experimental paradigm used in the research reported here, however, the authors of the above paper not only were concerned with the average response of subjects to a series of questions, but were also interested in the small quantum of additional insight that each subject could provide. It was only through a careful series of probing questions that this additional information was elicited. A fortunate additional result of the research reported here, then, is the further demonstration of the utility of performing well designed and controlled testing and observation of a system before putting that system into general use.

FUTURE STUDIES AND EXTENSION OF RESEARCH

The research reported here, in addition to providing useful insights into the display variables which affect the performance of the user in MMI, also establishes a methodology which may prove of use to future systems designers. The research began by attempting to develop a mathematical model of MMI. It is apparent that there would need to be man (user) oriented variables and machine oriented variables in the model in order to adequately describe user behavior. A systems designer, however, may be less interested in developing a general description of MMI and more interested in predicting user performance with a particular population, a particular set of values of the display variables, and a particular MMS. To the extent that the general model proposed here adequately describes user performance, and the performance measures used are of value to the designer, all that may be required is simply to solve the equations for performance, by fixing the parameters. There may still be concern, however, as to the adequacy of the performance measures used in this research. Also, if the target user population is not relatively homogeneous, then steps must be taken to ascertain those user-oriented variables (I.Q., general attitude towards computers, typing ability, specific computer or system experience, etc.) which also affect user performance. Heafner [1974], and Carlisle [1974], among others, postulate, or demonstrate, the effects that user characteristics have upon performance in interactive problem-solving tasks.

A number of the results presented in the previous chapter are not easily explained within the traditional confines of computer science and human factors research. The unusual pattern of interactions between baud rate and output variability upon the answers in questions 3, 7, 9, 14, 15 and 18 (see Figure 5-9) has been discussed in some detail in Chapter 5, but it is interesting to re-examine the results of the post-test questionnaire at this point as a means of pointing out the need for a broader theoretical base upon which to interpret the results. These six questions do not appear to have any immediately observable properties in common. In fact, they represent all four of the question groups discussed in Chapter 5: Command questions [C], Display questions [D], Speed questions [S] and the General Utility questions [U]. One of the most startling aspects of the responses to the post-test questionnaire is that only three questions, 10, 12 and 13, yielded results similar to what was conjectured before the study began: that there would be main effects for baud rate and output variability, with nonsignificant interaction effects. The conjectured results are presented below (Figure 6-1).

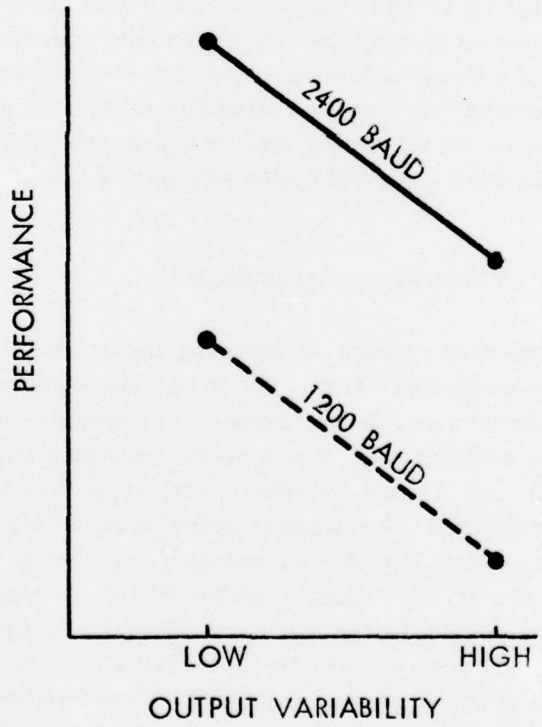


Figure 6-1. Prototypical results of post-test questionnaire

The deviations from the "expected" results generally took the form of non-significance of the effect for difference in display rate. Even on the [S] (speed and variability) group of questions (9-13), evaluated on a question-by-question basis, and on the effect of the display rate on the "index of satisfaction" (the average answer over the questions within the group) there was a surprising lack of baud rate effect. For the same "index of satisfaction" analysis, using the output variability as the independent variable, the effect was significant. Specifically, it would be of value to explore in greater depth the reasons why those subjects who received the slowest, most frustrating version of the system -- 1200 baud, high variability -- had responses to the questionnaire which in general were *higher* than those who received the 2400 baud, high variability version. Potential motivational reasoning was explored briefly in Chapter 5. It seems apparent that there ought to be a discipline of computer interaction, or computer programming, psychology which would provide a theoretical basis for a synthesis between cognitive psychology and computer science.

As indicated above, the results obtained in these experiments are generalizable to other users and other interactive systems only if we are willing to accept as reasonable the description of the population from which the subjects were sampled, and from which the Travel Message system is an example. Furthermore, it seems apparent that there are motivational and cognitive processes occurring which require an analysis outside the scope of this research. Though it may be reasonable to interpolate the results to values of the independent variables between the extremes tested here, the extrapolation to values outside the range tested is unjustified. As terminals become available, with faster display rates the conclusion of no effect on performance in increasing baud rates ought to be subjected to further scrutiny. If in fact increasing the display rate to 4800 baud, 9600 baud, or higher fails to produce significant user performance improvements, then system designers ought to be spending their resources in other areas. One area specifically identified in this research is the reduction of output rate variability, which can be decreased either by decreasing the nominal display rate or by increasing the power of the CPU. The research reported here supports the conclusion that within the limits of the variables studied here, if decreasing the nominal baud rate from 2400 to 1200 baud decreases the variability of the output, performance improves or is not decreased.

MISSING PAGE
NUMBERS ARE BLANK
AND WERE NOT
FILMED

BIBLIOGRAPHY

Ambrozy, Denise, "On Man-Computer Dialogue," in *Int. J. Man-Machine Studies*, Vol. 3, 1971, 375-383

Babbie, Earl R., *Survey Research Methods*, Wadsworth Publishing Co., Inc., Belmont, Ca., 1973

Bennett, John L., "The User Interface in Interactive Systems," in Cuadra, Carlos A. (ed.), *Annual Review of Information Science and Technology*, Vol. 7, 1972, ASIS, Washington, D.C.

Boies, S.J., "User Behavior on an Interactive Computer System," in *IBM Systems Journal*, No. 1, 1974, 2-19

Bolt Beranek and Newman, *Terminal Interface Message Processor User's Guide*, Report No. 2183, NIC No. 10916, Bolt Beranek and Newman, Inc., Cambridge, Mass., 1974

Burchfiel, Jerry D., Elsie M. Leavitt, Sonya Shapiro & Theodore Strollo, *TENEX Users' Guide*, Bolt Beranek and Newman, Inc., Cambridge, Mass., 1975

Campbell, Donald T. & Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research*, Rand McNally and Co., Chicago, 1963

Carlisle, James H., *Man Computer Interactive Problem Solving -- Relationships Between User Characteristics and Interface Complexity*, Ph.D. Thesis, Yale University School of Organization and Management, 1974

Cooper, William S., "On Selecting a Measure of Retrieval Effectiveness," in *J.ASIS*, March-April, 1973, 87-100

Cooper, William S., "On Selecting a Measure of Retrieval Effectiveness, Part II. Implementation of the Philosophy," in *J.ASIS*, Nov.-Dec., 1973, 413-424

Davies, D. R., & G. S. Tune, *Human Vigilance Performance*, American Elsevier Publishing Company, Inc., New York, 1969

De Greene, Kenyon B., "Man-Computer Interrelationships," in De Greene, Kenyon (ed.), *Systems Psychology*, McGraw Hill, New York, 1970, 281-336

Hansen, James V., "Man-Machine Communication: An Experimental Analysis of Heuristic Problem-Solving Under On-Line and Batch-Processing Conditions," in *IEEE Trans. Systems, Man, and Cybernetics*, Vol. SMC-6, No. 11, November, 1976, 746-752

Hansen, Wilfred J., "User engineering principles for interactive systems," in *FJCC* 1971, 523-532

Heafner, J. F., *A Methodology for Selecting and Refining Man-Computer Languages to Improve Users' Performance*, University of Southern California, Information Sciences Institute Research Report, ISI/RR-74-21, September, 1974

Heafner, J. F. & Lawrence Miller, *Design Considerations for a Computerized Message Service Based on Triservice Operations Personnel at CINCPAC Headquarter, Camp Smith, Oahu*, University of Southern California, Information Sciences Institute Working Paper, WP-3, September, 1976

Kerlinger, Fred N. & Elazar J. Pedhazur, *Multiple Regression in Behavioral Research*, Holt, Rinehart and Winston, Inc., New York, 1973

Mackworth, Jane F., *Vigilance and Attention: A Signal Detection Approach*, Penguin Books, Middlesex, England, 1970

Martin, Thomas H., James Carlisle, & Siegfried Treu, "The User Interface for Interactive Bibliographic Searching: An Analysis of the Attitudes of Nineteen Information Scientists," in *J.ASIS*, March-April, 1973, 142-147

Melnyk, Vera, "Man-Machine Interface: Frustration," in *J.ASIS*, Nov.-Dec., 1972, 392-401

Miller, Lance A., & Curtis A. Becker, *Programming in Natural English*, IBM Research Report, RC 5137, November, 1974

Miller, Robert B., "Response time in man-computer conversational transactions," in *American Federation of Information Processing Societies, Fall Joint Computer Conference, San Francisco, California, 1968, Proceedings*, Vol. 33, Part 1, Thompson Book Company, Washington, D.C., 1968, 267-277

Mostofsky, David I. (ed.), *Attention: Contemporary Theory and Analysis*, Appleton-Century-Crofts, New York, 1970

Nickerson, Raymond S., Jerome I. Elkind and Jaime R. Carbonell, "Human Factors and the Design of Time Sharing Computer Systems," in *Human Factors*, Vol. 10, No. 2, 1968, 127-134

Sackman, H., & Ronald L. Citrenbaum (eds.), "Human Factors Experimentation in Interactive Planning," in *ONLINE PLANNING, Towards Creative Problem Solving*, Prentice Hall, 1972

Salton, G., *Interactive Information Retrieval*, Cornell University, Department of Computer Science Technical Report No. 69-40, August, 1969,

Salton, G., "Dynamic Document Processing," in *C.ACM*, Vol. 15, No. 7, July, 1972, 658-668

Seven, M. J., B. W. Boehm & R. A. Watson, *A Study of User Behavior in Problem Solving with an Interactive Computer*, The Rand Corporation R-513-NASA, April, 1971

Shneiderman, Ben, "Exploratory Experiments in Programmer Behavior," in *International J. of Computer and Information Sciences*, Vol. 5, No. 2, 1976, 123-143

Smith, Robert, *Tenex SAIL*, Stanford University, Institute for Mathematical Studies in the Social Studies, Technical Report No. 248, January, 1975

Sterling, Theodor D., "Guidelines for Humanizing Computerized Information Systems: A Report from Stanley House," in *C.ACM*, Vol. 17, No. 11, November, 1974

Swets, John A., et al, *Information Processing Models and Computer Aids for Human Performance*, BBN Report No. 2008, Bolt Beranek and Newman, Inc., Cambridge, Mass., 31 July, 1970

Tomeski, Edward A. & Harold Lazarus, *People Oriented Computer Systems*, Van Nostrand Reinhold Company, New York, 1975

VanLehn, Kurt A. (ed.), *SAIL User Manual*, Stanford University Computer Science Department Report STAN-CS-73-373, July, 1973

Walker, Donald E. (ed.), *Interactive Bibliographic Search: The User/Computer Interface*, AFIPS Press, Montvale, N.J., 1971

Walther, George H. & Harold F. O'neil, Jr., "On-line user-computer interface--the effects of interface flexibility, terminal type, and experience on performance," in *National Computer Conference*, 1974

Willmorth, N. E., "Human Factors Experimentation in Interactive Planning," in Sackman, H. & Ronald L. Citrenbaum (eds.), *ONLINE PLANNING, Towards Creative Problem Solving*, Prentice Hall, 1972, 281-313

Winer, B. J., *Statistical Principles in Experimental Design*, McGraw-Hill, New York, 1971

APPENDIX I

INSTRUCTIONS TO SUBJECTS

You are about to participate in an experiment that is designed to help us produce interactive programs which are better and easier to use. In order to do this, we have selected an environment which allows us to observe a typical interaction session.

Not trying to disguise what we are doing, it is important to point out that this is a test of certain aspects of the system you will be using, and is in no way a test of individual performance. In a sense there are no right or wrong answers to the set of tasks and questions you will be asked to perform. In fact, you may assume that certain aspects of the system have been intentionally designed to be less than optimal, in order to determine whether they do in fact affect user performance. Different people in these experimental sessions will be performing their tasks with different versions of the system, and a comparison of the grouped data will be made.

You are to assume that you are a clerk in the travel department of a company. Individuals in the company make requests to the travel department for flights to various cities, using a computerized message creation and transmittal system. You utilize a modified version of the MSG system, renamed Travel Messages Processing System, to access the data base of travel requests.

You may assume that each request for a flight actually ended up in a flight, i.e., any cancellations of requests caused the initial request to be purged from the data base.

In your position as clerk in the travel department, you will be given a set of tasks relating to these travel requests. For example, who wanted to travel to N.Y. on such and such a date, etc. These tasks are answered by making searches through the data base, and reading the retrieved messages.

There are eleven tasks to be performed, plus two additional sample tasks at the beginning which will allow you to test your understanding of the system, experiment with the commands available, and determine what typical messages in the data base look like. Reference materials available for this session include the list of commands, the instructions for working the sample tasks and the experimenter for answering questions concerning the use of the system's functions and commands.

After you have completed the series of tasks, you will be asked to complete a questionnaire relating to your experiences with the system. It is important to understand that this questionnaire asks you to rate on a numerical scale certain features of the system. Please do no hesitate to give a negative rating, if you have negative feelings regarding a particular area. Similarly, do not hesitate to give a positive rating if you feel positively towards a question. Also, you must attempt to answer these questions based on your current use of the Travel Messages Processing System. Please do not answer based on your general knowledge of MSG or computers.

Since questionnaires tend to include areas where you may have no strong opinion, and exclude areas where you may have a strong opinion, there is a free-text input question at the end, which allows you to express your general comments on the system, including those areas which you feel were not adequately covered in the previous questions.

It is important to emphasize that your participation in this experiment is voluntary. You may withdraw from this experiment at any time. Though there may be no immediate benefits to you from this experiment, it is hoped that the results of this research may guide system designers in the future in producing interactive programs which are easier to use.

APPENDIX 2

TASKS TO BE PERFORMED

S1) HOW MANY REQUESTS WERE MADE TO THE TRAVEL DEPARTMENT for travel to San Diego? Follow the instructions on the answer sheet to answer this question.

S2) WHO WANTED TO GO TO DES MOINES DURING THE MONTH OF JANUARY? Again, follow instructions on the answer sheet to answer this question.

1) WHO WANTED TO GO TO LONDON IN MARCH? If there is more than one person who wanted to go, write down all of their names. If nobody wanted to travel to London in March, write "NONE."

2) WHO WANTED TO GO TO KANSAS CITY DURING THE MONTH OF FEB? Answer this question similarly to the previous question.

3) WHO WANTED TO GO TO PORTLAND DURING THE MONTH OF FEB? Answer this question similarly to the previous questions.

4) WHO WANTED TO GO TO MIAMI ON FEB. 2? Answer this question similarly to the previous questions.

5) WHO WANTED TO GO TO SAN DIEGO ON APRIL 2? Answer this question similarly to the previous questions.

6) HOW MANY REQUESTS FOR TRIPS TO SEATTLE ARE THERE IN THE DATA BASE?

7) WHO TOOK THOSE TRIPS, and how many trips did each of these people take to Seattle.

8) FOR THOSE WHO TOOK FIVE OR MORE TRIPS TO SEATTLE, to which other cities did they REQUEST travel?

9) LIST THE LOCATIONS AND REQUESTED DATES OF TRAVEL, that Alan Schwartz made, where he requested Sim Farar to also travel. Similarly, list locations and dates of travel where Sim Farar requested travel with Alan Schwartz.

10) List the dates when Alan Schwartz and Sim Farar both traveled together to Seattle.

11) ON THOSE OCCASIONS WHERE BOTH SIM FARAR AND ALAN SCHWARTZ TRAVELED TOGETHER TO SEATTLE, list those who also traveled with them.

APPENDIX 3

POST-TEST QUESTIONNAIRE

INSTRUCTIONS FOR POST-TEST QUESTIONNAIRE

You are now requested to answer a brief series of questions concerning your opinions of the computerized system you've just been using. It is important that you answer these questions with the answer that best represents your attitude to the particular area of the question. Specifically, the questions require you to numerically rate certain aspects of the computerized system. Even though you may not have a strong feeling one way or the other, please select one of the numerical ratings that best characterizes your attitude to that particular area.

You will note that the questions are answered using the computer. Please be careful that you select the correct number for your answer. If you make an error, press the "DEL" (or A) key. When you are satisfied with your answer for that particular question, press the "RETURN" key.

> Please answer the following series of questions with a numerical rating in the range of 1-5. 1 = Very Poor, Unacceptable, etc. 5 = Excellent, Completely Acceptable, Easy to Use, etc. (1-2 implies a generally negative response, 4-5 a generally positive one.) However, a specific numerical scale will be given for each question.

1) COMMANDS: EASE OF USE--

1 = Difficult to use.
3 = Easy to use, but somewhat confusing.
5 = Easy to use, no confusion as to meaning.

2) COMMANDS: CLEAR AND MEANINGFUL FUNCTIONS--

1 = Commands produced results completely different from what was expected.
3 = Some commands were clear and simple, others were very confusing.
5 = All commands were completely clear.

3) COMMANDS--

1 = Would liked to have had a number of additional commands available to make the tasks easier to accomplish.
3 = Some additional commands would have been useful.
5 = Available commands were completely adequate to accomplish tasks.

4) SCREEN: BRIGHT ENOUGH?

1 = Too dim, completely unreadable.
3 = Too dim, but readable.
5 = Brightness just right.

5) SCREEN: LARGE ENOUGH?

1 = Screen size too small, completely unreadable
3 = Screen size too small, but readable
5 = Screen size just right

6) CHARACTERS: LEGIBLE, ADEQUATE SIZE, ETC.--

1 = Characters too small or awkwardly shaped
3 = Character size and shape adequate, but some difficulty in reading
5 = Character size and shape just right

7) PRINTING FORMAT: READABLE?

1 = Format unclear, jumbled, etc. Unreadable.
3 = Format readable, but not outstanding.
5 = Format excellently arranged and completely readable.

8) PRINTING FORMAT: SUFFICIENT DATA?

1 = Completely insufficient data to adequately complete tasks.
3 = Just barely sufficient data, but would have been able to utilize more.
5 = Data presented was completely adequate to complete tasks.

9) COMPUTER SYSTEM SPEED--

1 = Too slow
3 = Just right
5 = Too fast

10) VARIATION IN COMPUTER SYSTEM SPEED--

1 = So much variation in computer and printing speed that system was difficult and bothersome.
3 = Some variation in computer speed and printing speed, but not enough to be bothersome.
5 = Little or no variation in the speed of the computer system.

11) PRINTING SPEED--

1 = Too slow
3 = Just right
5 = Too fast

12) VARIATION IN PRINTING SPEED--

1 = Far too much variation for easy reading of output
3 = Some variation, but no great difficulty in reading
5 = Output was smooth and easy to read

13) PROCESSING TIME--

1 = System took way too long to do what should have been simple tasks.
3 = System took about the time you would have expected.

5 = Too fast, felt rushed, etc.

14) WAS THE TRAVEL MESSAGE PROCESSING SYSTEM USEFUL IN ANSWERING THESE QUESTIONS?

1 = Completely useless, confusing, etc. Answering the questions was an exercise in futility.
3 = Found the system marginally useful, but some aspects were difficult to use, too slow, confusing, etc.
5 = Completely useful, no confusion in the use of the system. Speed of system was just right, easy to adapt to.

15) SUPPOSE YOU HAD TO ACTUALLY ANSWER THE TRAVEL QUESTIONS BY GOING THROUGH THE MESSAGES BY HAND. HOW MUCH IS THE TRAVEL MESSAGES COMPUTER PROCESSING SYSTEM WORTH TO YOU IN ORDER TO SAVE YOU THE EFFORT OF DOING THIS BY HAND?

1 = No advantage seen in using the computer system. Would much prefer to perform these tasks by hand.
3 = No strong feeling one way or the other.
5 = Much prefer using the computer system rather than having to answer these questions by going through the messages by hand.

16) DID YOU FEEL A NEED FOR MORE MATERIALS ON THE FUNCTIONS AVAILABLE IN THE SYSTEM?

1 = Available materials were completely useless.
3 = What was available was useful, but more information was needed.
5 = All available material was useful, no more information was needed.

17) YOUR OVERALL RATING OF INPUT TO THE COMPUTER--
[Use a 1-5 scale as explained in the top portion of the screen.]

18) YOUR OVERALL RATING OF OUTPUT FROM THE COMPUTER--
[Use a 1-5 scale as explained in the top portion of the screen.]

Please type your general comments on the functions provided, their ease of use, and your general feelings of frustration or satisfaction in the use of the system. Be certain to address yourself to your feelings in regards to the delays in output, and the general speed of the system, particularly if the load average was high and you noted unacceptable delays in system performance.

APPENDIX 4

SAMPLE MESSAGES

To: TRAVEL DEPT.
From: Alan Schwartz
Subject: Travel, San Francisco, Feb. 2 a.m.
Date: 31 JAN 76 1303-PST
Message:
Please reserve 2 seats to San Francisco on Feb. 2 a.m. for me and
    Arnold Serkin
Return: OPEN
Thanks

To: TRAVEL DEPT.
From: David Simpson
Subject: Travel, Des Moines, Jan. 4 p.m.
Date: 31 JAN 76 1303-PST
Message:
Please reserve 4 seats to Des Moines on Jan. 4 p.m. for me and
    Jane Doe
    Arnold Serkin
    John Wilson
Return: Jan. 8
Thanks

To: TRAVEL DEPT.
From: Arnold Serkin
Subject: Travel, Miami, April 23 a.m.
Date: 31 JAN 76 1303-PST
Message:
Please reserve 3 seats to Miami on April 23 a.m. for me and
    Sim Farar
    Alan Schwartz
Return: April 27
Thanks

To: TRAVEL DEPT.
From: Larry Miller
Subject: Travel, San Diego, April 2 a.m.
Date: 31 JAN 76 1303-PST
Message:
Please reserve 3 seats to San Diego on April 2 a.m. for me and
    David Simpson
    John Wilson
Return: April 6
Thanks

DISTRIBUTION LIST

Defense Advanced Research Projects Agency
1400 Wilson Blvd., Arlington, VA 22209

Defense Documentation Center
Cameron Station
Alexandria, VA 22314

Steven Swart,ACO
Office of Naval Research
1030 Green Street
Pasadena, CA 91106